

VISIT...

LANZAROTE
Caliente.COM

Phenomenal Concepts and Phenomenal Knowledge

New Essays on Consciousness and Physicalism

Alter, Torin Associate Professor of Philosophy, University of Alabama

Walter, Sven Junior Lecturer, Department of Philosophy, Universität Bielefeld

Contents

Introduction 1

Part One - Phenomenal Knowledge 13

1. Daniel Dennett - What RoboMary Knows 14
2. Laurence Nemirow et. altri - So This Is What It's Like A Defense of the Ability Hypothesis 32
3. Frank Jackson - The Knowledge Argument, Diaphanousness, Representationalism 52
4. Torin Atler - Does Representationalism Undermine the Knowledge Argument? 65
5. Knut Nordby - What Is This Thing You Call Color. Can a Totally Color-Blind Person Know about Color?

Part Two - Phenomenal Concepts 85

6. Janet Levin - What Is a Phenomenal Concept? 87
7. David Papineau - Phenomenal and Perceptual Concepts 111
8. Joseph Levine - Phenomenal Concepts and the Materialist Constraint 145
9. David Chalmers - Phenomenal Concepts and the Explanatory Gap 167
10. John Hawthorne - Direct Reference and Dancing Qualia 195
11. Stephen White - Property Dualism, Phenomenal Concepts, and the Semantic Premise 210
12. Ned Block - Max Black's Objection to Mind-Body Identity 249
13. Martine Nida-Rümelin - Grasping Phenomenal Properties 307

Introduction

This volume presents thirteen new essays on phenomenal concepts and phenomenal knowledge: twelve by philosophers and one by a scientist. In this introduction, we provide some background and summarize the essays.

Background: Consciousness and Physicalism

“Phenomenal” indicates conscious experience: what it's like to feel pain, to see red, and so on. *Phenomenal knowledge* is knowledge of conscious experience. *Phenomenal concepts* are concepts associated with that knowledge: those that express phenomenal

qualities from the experiencing subject's perspective. What is the nature of such knowledge and concepts? What are their distinctive characteristics? How are their contents determined? What is required for their possession? In what does possessing them consist? How are they related to abilities, such as the ability to visualize? How are they related to physical **knowledge** and physical concepts? These are just a few of the questions that the essays in this volume address.

Why are such questions important? One reason concerns the debate over consciousness and physicalism. *Physicalism* (also known as *materialism*) is the view that the world is merely physical.¹ This view implies that consciousness is physical or, as it is sometimes put, that there are no truths about consciousness over and above the physical truths. Many embrace this implication, partly because it accords well with a naturalistic outlook in which the physical sciences, ideally conceived, completely describe reality. But the implication is controversial even among those otherwise sympathetic to naturalism. The controversy has gained focus over the last few decades, partly because of the refinement of certain antiphysicalist arguments and physicalist replies. In this debate, phenomenal concepts and phenomenal knowledge have come to play increasingly prominent roles. To see why, consider two widely discussed antiphysicalist arguments: the knowledge argument and the conceivability argument.

The classic statement of the knowledge argument comes from Frank Jackson (1982, 1986). Jackson formulates the argument in terms of his well-known thought experiment: the case of Mary the super-scientist. Mary is raised in a black-and-white room and has no color experiences. She learns everything in the completed science of color vision by watching lectures on black-and-white television. What she learns includes “everything in completed physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course functional roles” (Jackson 1986: 291). Then she leaves the room or is given a color television.

Jackson uses the Mary case to argue roughly as follows. Before she leaves the room, she knows all the physical truths. Intuitively, when she leaves and sees colors for the first time, she learns new truths about what it's like to see in color, truths she cannot deduce from those conveyed by black-and-white lectures. These new truths must be nonphysical; otherwise, she would have known them before leaving the room. If such nonphysical truths exist, then physicalism is false. Therefore, physicalism is false. That, in brief, is the knowledge argument.²

The conceivability argument is often formulated in terms of another thought experiment, one involving zombies. Zombies are defined as creatures that lack phenomenal consciousness but are physically identical to conscious human beings. Given this definition, the conceivability argument runs roughly as follows. Intuitively, zombies are conceivable. Their conceivability does not derive from our ignorance or cognitive limitations; rather, they are conceivable because they are metaphysically possible. If they are metaphysically possible, then physicalism is false. Therefore, physicalism is false.³

Both arguments comprise three main steps. The first step is to establish an *epistemic gap*: the thesis that phenomenal concepts/knowledge cannot be deduced a priori from physical concepts/knowledge. In terms of the knowledge argument, the claim is that Mary gains factual knowledge when she leaves the room: knowledge of truths that cannot be a priori deduced from her comprehensive physical knowledge. In terms of the conceivability argument, the claim is that it is possible to form a coherent positive conception of

zombies or that no a priori reasoning could show the zombie hypothesis to be incoherent (Chalmers 2002). The second step is to infer a *metaphysical gap* from the epistemic gap: to argue that the epistemic gap reflects a gap in reality, between the physical and the phenomenal themselves, and not just in knowledge or concepts. In terms of the knowledge argument, the inference is to the claim that the truths Mary learns upon leaving the room are nonphysical. In terms of the conceivability argument, the inference is to the claim that zombies are not just conceivable but metaphysically possible. The third step is to infer physicalism's falsity from the existence of a metaphysical gap. Most physicalists accept the third step but reject the first or the second.⁴ Against the first, some contend that the appearance of an epistemic gap derives from misconstruing phenomenal concepts or phenomenal knowledge. In their view, the appearance of such a gap does not reflect reality: on reflection, zombies are inconceivable and Mary gains no factual knowledge upon leaving her room. Against the second step, some argue that the epistemic gap does not entail a metaphysical gap. On their view, zombies are conceivable but metaphysically impossible and the thesis that Mary acquires factual knowledge when she leaves the room does not conflict with physicalism. These two views—the one that rejects the epistemic gap and the one that rejects the inference to the metaphysical gap—are sometimes called *type-A* and *type-B materialism*, respectively (Chalmers 1996, 2003). Both views have been developed in various ways, and antiphysicalists have responded in kind. But many on all sides of the debate would agree that much depends on how phenomenal concepts and phenomenal knowledge are construed. The latter issue has implications for what form physicalism should take: certain construals render the epistemic gap implausible and thus favor type-A materialism, while other construals help provide grounds for rejecting the inference to the metaphysical gap and thus favor type-B materialism. Moreover, the issue of how to construe phenomenal concepts and phenomenal knowledge has implications for the antiphysicalist arguments. For example, some type-A materialists argue that the knowledge argument goes awry in assuming that phenomenal knowledge is a species of factual/propositional knowledge, and some type-B materialists argue that phenomenal concepts have distinctive features that explain why the conceivability of zombies fails to support their metaphysical possibility. Indeed, views about phenomenal concepts or phenomenal knowledge play pivotal roles in virtually all serious discussions of the antiphysicalist arguments.

Thus, two questions emerge:

1. Could a proper understanding of phenomenal concepts/knowledge show that there is or is not an epistemic gap?
2. Could a proper understanding of phenomenal concepts/knowledge show that there is or is not a metaphysical gap?

Most of the essays in this volume address at least one of these questions, and all address surrounding issues. The essays in part I focus primarily on phenomenal knowledge and the knowledge argument. The essays in part II focus primarily on phenomenal concepts, and most do not concentrate narrowly on any single antiphysicalist argument.

Part I: Phenomenal Knowledge

Part I begins with an essay by Daniel Dennett that further develops a line of argument he presented in his 1991 book, *Consciousness Explained*. In that book, he argued that we should reject the intuition that Mary gains knowledge when she leaves the room. In his view, this intuition derives from a failure to appreciate the implications of knowing *all* the physical facts. In chapter 1 of the present volume, he gives a more detailed account of his case. Specifically, he (1) criticizes attempts to defend the intuition; (2) devises variations on the Mary case to illustrate how a deduction from physical information of what it's like to see in color might actually proceed; and (3) defends his arguments against objections. In effect, he answers question 1 of the previous section ("Could a proper understanding of phenomenal concepts/knowledge show that there is or is not an epistemic gap?") affirmatively. In his view, a proper understanding of phenomenal concepts and phenomenal knowledge helps to show that there is no epistemic gap.

In chapter 2, Laurence Nemirow also provides grounds for denying the existence of an epistemic gap. Unlike Dennett, however, Nemirow accepts that Mary learns what it's like to see in color when she leaves the room. But in his view, her learning consists in acquiring abilities, such as the ability to imagine seeing red, as opposed to phenomenal information. David Lewis (1988) dubbed the idea that knowing what it's like is possessing abilities *the ability hypothesis*, and Nemirow (1980, 1990) was its pioneer. In his chapter for this book, Nemirow defends the ability hypothesis and its effectiveness in undermining the knowledge argument. He considers a variety of objections, including objections advanced by Earl Conee, Michael Tye, Janet Levin, Brian Loar, Martine Nida-Rümelin, William Lycan, Torin Alter, and John Perry. As he mentions, many philosophers find the ability hypothesis counterintuitive. But in his view, it should be judged by "the strength of the available rejoinders," and on that score, he argues, it "proves to be reasonably resilient to assault."

Jackson's current position on the knowledge argument is not far from Nemirow's. In the late 1990s, Jackson surprised the philosophical community by embracing physicalism and rejecting the knowledge argument. In a 2003 essay, "Mind and Illusion," he argued that the claim that Mary learns new truths when she leaves the room—a version of the epistemic gap—derives from a mistaken conception of sensory experience, one that should be replaced with *representationalism*, the view that phenomenal states are representational states. In chapter 3, he further develops his representationalist view about perceptual experience and defends its application to the knowledge argument. He bases his view partly on the idea that perceptual experience is diaphanous—in other words, that "accessing the nature of the experience itself is nothing other than accessing the properties of its object." He argues that although the diaphanousness thesis alone does not entail representationalism, the thesis supports an inference from a weaker to a stronger version of representationalism. On the weak version, perceptual experience is essentially representational. On the strong version, "how an experience represents things as being exhausts its experiential nature." And strong representationalism, he argues, undermines the claim that Mary learns new truths when she leaves the room.

In chapter 4, Torin Alter criticizes the main argument of Jackson's "Mind and Illusion." In effect, Alter defends the epistemic gap against Jackson's argument from

representationalism. He argues that it is possible to formulate a representationalist version of the knowledge argument that inherits the force of the original. He concludes on these grounds that “in the debate over the knowledge argument, representationalism would appear to be a red herring.” He thus defends a version of the epistemic gap.⁵

The final chapter in part I, by the late Knut Nordby, takes a different approach. Nordby was a real-life counterpart of Mary: a color-blind expert in the science of color vision. In chapter 5, he describes the results of empirical research on color vision and other sense modalities. Based on these results and his own experience, he argues that “Mary will be able to sense and discriminate color hues, but will not be able to name them on the basis of her knowledge.” He does not take a definite stand on the epistemic and metaphysical gaps. But his reflections should help inform views on these matters.⁶

Part II: Phenomenal Concepts

Most of the essays in part II concern the view that, although consciousness is physical, our concepts of consciousness are special in ways that our concepts of other physical phenomena are not. On this view, the antiphysicalist arguments mistake an insight about phenomenal concepts for one about what these concepts pick out. On the most popular version of the view, known as *the phenomenal concept strategy* (Stoljar 2005), the epistemic gap is accepted but explained in psychological terms.⁷ This strategy requires an account of phenomenal concepts that both explains the epistemic gap and is consistent with physicalism. The first two essays in part II (by Janet Levin and David Papineau) defend accounts of phenomenal concepts meant to meet both desiderata. The next two essays (by Joseph Levine and David Chalmers) criticize the phenomenal concept strategy. They are followed by a chapter in which John Hawthorne criticizes the combination of antiphysicalism and a view about phenomenal concepts typically held by antiphysicalists. The last three (by Stephen White, Ned Block, and Martine Nida-Rümelin) discuss antiphysicalist arguments by way of issues that center on phenomenal concepts.

In chapter 6, Levin presents a version of the phenomenal concept strategy based on a limited defense of the “demonstrative account” of phenomenal concepts. In this account, phenomenal concepts “pick out their referents *directly*, much like demonstratives, without mediation by any mode of presentation.” As she observes, many type-B materialists appeal to this account to help explain why there is an epistemic gap but no metaphysical gap. She argues that the account can meet objections she and others present elsewhere, although she notes that other objections remain unanswered. She also argues that recent emendations to the account, including those by Katalin Balog, Block, Papineau, and Levine, “concede too much to the antiphysicalists while accomplishing too little.” She therefore urges demonstrative theorists “to return to their roots.”

Although Papineau once advocated a demonstrative account, in his recent book, *Thinking about Consciousness* (2002), he argues that phenomenal concepts should instead be likened to quotational terms: such concepts are, he suggests, “terms with the structure *the experience*: —, in which the gap is filled either by a current experience or by an imaginative re-creation of an experience.” The quotational account, like the

demonstrative account, can be used to defend type-B materialism from the antiphysicalist arguments. For example, he argues, although leaving the room allows Mary to fill the placeholder in “the experience: —” with the appropriate color experiences, her new phenomenal concepts are merely distinctive ways of referring to physical phenomena. In chapter 7, he develops a revised version of the quotational account. The revisions are motivated in part by the need to explain how one who possesses a phenomenal-red concept, for example, could think truly that she is now neither having a red experience nor re-creating such an experience in her imagination. Although the revisions are substantial, he argues that “the main arguments in the book [*Thinking about Consciousness*]”—including the type-B-style responses to the antiphysicalist arguments—“are robust with respect to” these revisions.

In chapter 8, Levine raises a problem for the phenomenal concept strategy. He frames the problem partly in terms of the *explanatory gap* (Levine 1983), which is roughly the claim that the existence or nature of phenomenal consciousness cannot be completely explained in physical terms. The explanatory and epistemic gaps are closely related. Note, for example, that if zombies are conceivable, then the complete physical explanation does not appear to explain why consciousness exists—why our world is not a zombie world (see Chalmers, chap. 9, this volume). As applied to the explanatory gap, the phenomenal concept strategy requires a physicalist account of phenomenal concepts on which the gap derives from phenomenal concepts rather than phenomenal consciousness itself. Levine argues that, to pass muster, such accounts must satisfy the following constraint: “that no appeal be made in the explanation to any mental property or relation that is basic.” An account violates this constraint if, for example, it makes appeal to an unexplained notion of acquaintance between a subject and her brain states. He argues that we do not understand how any physicalist account can both meet this constraint and explain how the explanatory gap derives from “the peculiar features of phenomenal concepts.” And though he allows that some physicalist account might achieve these goals, he ends by speculating that physicalism may be false “not because phenomenal properties themselves are not physical” but rather “because somehow we embody a relation to them that is itself brute and irreducible to physical relations.”

In chapter 9, Chalmers raises a related problem for the phenomenal concept strategy. He argues that “no account of phenomenal concepts is both powerful enough to explain our epistemic situation with regard to consciousness, and tame enough to be explained in physical terms.” In other words, any account of phenomenal concepts that renders them physically explicable fails to capture our epistemic situation; and any account that captures our epistemic situation entails that phenomenal concepts are not physically explicable. Chalmers illustrates the problem by applying the dilemma to demonstrative accounts such as Levin's, quotational accounts such as Papineau's, and various other accounts. If he is right, then the problem is insurmountable, and the phenomenal concept strategy cannot succeed.

Antiphysicalists have problems of their own. In chapter 10, Hawthorne argues that there is a tension in the semantic views held by certain antiphysicalists. These philosophers accept Fregean arguments against direct-reference theories of ordinary proper names but maintain that phenomenal concepts refer directly. Against this semantic package, and against Chalmers's views in particular, Hawthorne argues that “the thought experiments that motivate a sense-reference distinction for ordinary proper names—roughly,

Hesperus-Phosphorus stories—can be replicated at the level of direct phenomenal concepts.” (A *Hesperus-Phosphorus story* is one in which one rationally believes both that object a has a property P and that object b lacks P, even though $a = b$.)⁸ Further, he suggests, “the thought that Hesperus-Phosphorus phenomena arise for direct phenomenal concepts” arises independently of the semantic package and thus “may have a quite general legitimacy.” Although he does not take a definite position on the existence of the epistemic and metaphysical gaps, in effect he identifies a new problem for those who embrace them.

The next two essays focus on what is known as *the property dualism argument*. Though the knowledge argument and the conceivability argument apply to physicalism generally, the property dualism argument targets a specific version: the identity theory, on which experiences are held to be identical to more fundamental physical phenomena, such as brain processes.⁹ The property dualism argument exploits a consideration related to the Hesperus-Phosphorus phenomena that Hawthorne discusses. The identity theorist accepts that one might rationally both believe that one is in pain and disbelieve that brain state B is occurring, even if $\text{pain} = B$. This is presumably because the subject thinks of one and the same event under two distinct “modes of presentation”: a phenomenal mode and a neurobiological mode. But how are we to understand phenomenal modes of presentation? According to the property dualism argument, a proper understanding of phenomenal modes leads to a dualism of properties.

In chapter 11, White defends the property dualism argument. The “semantic premise” mentioned in his chapter title refers to an assumption identified by Brian Loar (1990/97). According to Loar, antiphysicalist arguments such as the property dualism argument tacitly assume that “a statement of property identity that links conceptually independent concepts is true only if at least one concept picks out the property it refers to by connoting a contingent property of that property” (600). According to White, however, the “property that does the work in explaining the possibility of a posteriori identities needn’t be a first-order property of the referent in question.” On his view, the property dualism argument requires only a weaker semantic premise, which allows that the property in question be a higher order property. He formulates a refined version of the property dualism argument, which uses the weaker premise, and defends the argument against various objections.

In chapter 12, Block criticizes the property dualism argument. He argues that one version of the argument conflates two different notions of mode of presentation: the “cognitive mode of presentation,” which is defined in terms of its role in determining reference and/or explaining cognitive significance; and the “metaphysical mode of presentation,” which is a property of the referent in virtue of which the cognitive mode of presentation plays its semantic and cognitive roles. Block also criticizes other versions of the property dualism argument, and he draws connections to the knowledge argument and to related arguments given by White. In effect, Block and White disagree over the inference from the epistemic to the metaphysical gap: Block uses a version of the phenomenal concept strategy to reject the inference, whereas White defends the inference against that strategy. In the final chapter, Nida-Rümelin presents a different argument for property dualism. Central to her argument is a technical notion of *grasping a property*, which means “understand[ing] what having that property essentially consists in.” She develops this notion and uses it to formulate four theses about phenomenal concepts, physical concepts,

and the relations among these concepts and phenomenal and physical properties—theses that, she argues, together establish that “no phenomenal property can be a physical property.” She frames her discussion partly in terms of the two-dimensional semantic framework developed by Jackson, Chalmers, and others. At the end of her essay, she compares and contrasts her argument with related antiphysicalist arguments given by Chalmers (1996, 2003) and Kripke (1972).

Properly assessing the antiphysicalist arguments requires resolving various issues concerning phenomenal concepts and phenomenal knowledge. The essays collected here advance the debate about these issues and, we believe, thereby improve our understanding of the nature of consciousness and its relation to the physical world.

Acknowledgments

We are indebted to David Chalmers for his generous and invaluable advice on countless matters. We also wish to thank Norma McLemore, Karissa Rinas, and Mark Scala for their help with copy editing. Finally, Torin Alter thanks the American Philosophical Society for supporting his work on this book with a Sabbatical Fellowship.

References

- Campbell, K. (1970). *Body and Mind*. New York: Doubleday.
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Chalmers, D. J. (2002). Does Conceivability Entail Possibility? In *Conceivability and Possibility*, ed. T. Gendler and J. Hawthorne: 145–200. New York: Oxford University Press.
- Chalmers, D. J. (2003). Consciousness and Its Place in Nature. In *The Blackwell Guide to the Philosophy of Mind*, ed. P. Stich and T. Warfield. Oxford: Blackwell. Reprinted in *The Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 247–72. New York: Oxford University Press, 2002.
- Chalmers, D.J. (2004). Phenomenal Concepts and the Knowledge Argument. In *There's Something about Mary*, ed. P. Ludlow, D. Stoljar, and Y. Nagasawa: 269–98. Cambridge: MIT Press.
- Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown.
- Descartes, R. (1641). *Meditations on First Philosophy*.
- Dretske, F. (1995). *Naturalizing the Mind*. Cambridge: MIT Press.
- Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100: 25–50. Translated as On Sinn and Bedeutung, in *The Frege Reader*, ed. M. Beaney: 151–71. Oxford: Blackwell, 1997.
- Harman, G. (1990). The Intrinsic Quality of Experience. In *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 31–52. Atascadero, CA: Ridgeview.
- Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32, 127–36.



Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy* 83: 291–95.

 [Link ▶](#)

Jackson, F. (1998). Postscript on Qualia. In *Mind, Method, and Conditionals: Selected Essays*: 76–79. London: Routledge.

Jackson, F. (2003). Mind and Illusion. In *Minds and Persons: Royal Institute of Philosophy Supplement* 53, ed. A. O'Hear: 251–71. Cambridge: Cambridge University Press.

Kirk, R. (1974). Zombies versus Materialists. *Aristotelian Society Supplements* 48: 135–52.

Kripke, S. (1972). Naming and Necessity. In *The Semantics of Natural Language*, ed. G. Harman and D. Davidson. Dordrecht: Reidel. Reprinted as *Naming and Necessity*.

Cambridge: Harvard University Press, 1980.

Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64: 354–61.

Lewis, D. (1988). What Experience Teaches. *Proceedings of Russellian Society (University of Sydney)*. Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 281–94. New York: Oxford University Press, 2002.

Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives* 4: *Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, California: Ridgeview Publishing Co. Revised version in *The Nature of Consciousness*, ed. by N. Block, O. Flanagan, and G. Güzeldere: 597–616. Cambridge: MIT Press, 1997.

Nemirow, L. (1980). Review of *Mortal Questions*, by Thomas Nagel. *Philosophical Review* 89: 473–77.  [Link ▶](#)

Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. In *Mind and Cognition*, ed. W. Lycan: 490–99. Cambridge: Basil Blackwell.

Papineau, D. (2002). *Thinking about Consciousness*. New York: Oxford University Press.  [Link ▶](#)  [OSO X-Reference](#)

Putnam, H. (1967). Psychological Predicates. In *Art, Mind, and Religion*, ed. W. Capitan, and D. Merrill: 37–48. Pittsburgh: University of Pittsburgh Press. Reprinted as The Nature of Mental States, in Putnam, *Mind, Language and Reality*: 429–40, Cambridge: Cambridge University Press, 1975.

Stoljar, D. (2005). Physicalism and Phenomenal Concepts. *Mind and Language* 20: 469–94.  [Link ▶](#)

Stoljar, D., and Nagasawa, Y. (2004). Introduction to *There's Something about Mary*, ed. P. Ludlow, D. Stoljar, and Y. Nagasawa: 1–36. Cambridge: MIT Press.

Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge: MIT Press.

Part One Phenomenal Knowledge

end p.13

end p.14

one What RoboMary Knows

Daniel Dennett

Frank Jackson's thought experiment about Mary the color scientist is a prime example of an intuition pump, a thought experiment that is not so much a formal argument as a vignette that has been pumping philosophical intuitions with remarkable vigor since it first appeared in 1982. For sheer volume and reliability, this must count as one of the most successful intuition pumps ever devised by analytical philosophers. But is it a good intuition pump? How could we tell? Douglas Hofstadter's (1981) classic advice to philosophers confronted by a thought experiment is to treat it the way scientists treat a phenomenon of interest: vary it, turn it over, examine it from all angles, and in different settings and conditions, just to make sure you aren't taken in by illusions of causation. During the last twenty years, philosophers have examined many variations and defended many different responses, but they have singularly neglected some of the possible settings of the knobs. More than a decade ago, I conducted a preliminary exploration of the knobs, and issued a killjoy verdict that has been almost universally disregarded: "Like a good thought experiment, its point is immediately evident even to the uninitiated. In fact it is a bad thought experiment, an intuition pump that actually encourages us to misunderstand its premises!" (1991: 398).

In fact, it is much more difficult to imagine the scenario correctly than people suppose, so they imagine something easier and draw their conclusions from that mistaken base. In an attempt to bring out the flaws in the thought experiment, I encouraged people to consider a variant ending:

And so, one day, Mary's captors decided it was time for her to see colors. As a trick, they prepared a bright blue banana to present as her first color experience ever. Mary took one look at it and said, "Hey! You tried to trick me! Bananas are yellow, but this one is blue!" Her captors were dumfounded. How did she do it? "Simple," she replied. "You have to remember that I know *everything*—absolutely everything—that could ever be known about the physical causes and effects of color vision. So of course before you brought the banana in, I had already written down, in exquisite detail, exactly what physical impression a yellow object or a blue object (or a green object, etc.) would make
end p.15

on my nervous system. So I already knew exactly what *thoughts* I would have (because, after all, the 'mere disposition' to think about this or that is not one of your famous qualia, is it?). I was not in the slightest surprised by my experience of blue (what surprised me was that you would try such a second-rate trick on me). I realize it is *hard for you to imagine* that I could know so much about my reactive dispositions that the way blue affected me came as no surprise. Of course it's hard for you to imagine. It's hard for anyone to imagine the consequences of someone knowing absolutely everything physical about anything!" (1991: 399–400)

It is standardly assumed that things could not proceed this way. As Jackson disarmingly put it in the original article, "It seems just obvious that she will learn something about the world and our visual experience of it" (1982: 128). That, I claimed, is a mistake, and that

is what is wrong with Mary as a thought experiment. It just feels so good to conclude that Mary has a revelation of *some* sort when she first sees color that nobody wants to bother to show that this is how the story must go. In fact, it needn't go that way at all. My variant was intended to bring out the fact that, absent any persuasive argument that this could not be how Mary would respond, my telling of the tale had the same status as Jackson's: two little fantasies pulling in opposite directions, neither with any demonstrated authority. I thought that I had said enough to make my point, but a decade of further writing on Mary by many philosophers and their students has shown me that I should have been more patient, more explicit, in my objections. I underestimated the allure of this intuition pump by a wide margin. So I am returning to the fray, and this time I will make my case at a more deliberate pace, dotting the i's and crossing the t's.

First, I have found that some readers—maybe most—just didn't get my blue banana alternative.¹ What was I saying? I was saying that Mary had figured out, using her vast knowledge of color science, *exactly what it would be like for her to see* something red, something yellow, something blue in advance of having those experiences.² I asserted this flat out—in your face, as it were—to expose the fact that people normally assume that this is impossible on the basis of no evidence or theory or argument, but just on the basis of ancient philosophical tradition going back at least to John Locke. Perhaps a little dialogue will help bring out the intended point:

TRAD: What on earth do you mean? *How* could Mary do that?

DCD: It wasn't easy. She deduced it, actually, in a 4,765-step proof (for red—once she'd deduced what red would look like to her, green fell into line with a 300-step lemma, and the other colors, and all the hues thereof, were relatively trivial extensions of those proofs).

end p.16

TRAD: You're just making all that up! There are no such proofs!

DCD: This is a thought experiment; I get to make up all sorts of things. Can you prove that there are no such proofs? What established fact or principle am I contradicting when I help myself to a scenario in which she deduces what colors would look like to her from everything she knows about color?

TRAD: Look. It's just obvious! *You can't deduce what a color looks like if you've never seen one!*

DCD: That's an interesting folk theorem, I must say. Here's another: If you burp, sneeze, and fart all at the same time, you die. Both sound sort of plausible to me, but is there any scientific backing for either one of them?

She'll Be Surprised, Dammit!

If the Mary thought experiment was intended simply to draw out and illustrate vividly the implications of a fairly standard way of thinking that many, probably most, people have, it might be a useful anthropological exercise, an investigation of folk psychology laid bare. But those who have championed Mary have thought that it might actually prove something bigger, not just the conclusion that most people's unexamined assumptions imply dualism (I think we already knew that, but maybe not), but the conclusion that dualism is true! The fact that philosophers would so much as *entertain* such an interpretation of a casual exercise of the imagination fills me with astonishment. I had no

idea philosophers still put so much faith in the authority of their homegrown intuitions. It is almost as if one thought one could prove that the Copernican theory was false by noting that it “seems obvious” that the earth doesn't move and the sun does.

Consider, for instance, the recent article “Mary Mary, Quite Contrary,” by George Graham and Terence Horgan (2000). Graham and Horgan have usefully managed to distill precisely the unargued intuition that I have been attempting to isolate and discredit for fifteen years or more—the one we might express as “She'll be surprised, dammit!”

They begin by distinguishing two main materialist responses to Mary: thin and thick materialism. Thin materialism, of which I am one of the few exponents, denies that Mary learns anything post-release. Thick materialists attempt to salvage materialism while going along with the gag that Mary is startled, delighted, surprised, or something like that, when she is released from her colorless captivity. The strategy Graham and Horgan take is first to declare briskly that thin materialism is a nonstarter in need of no refutation because it “has been amply criticized by others” (63). The only critics they list are McConnell (1994) and Lycan (1996). As I replied at some length to McConnell in the same journal (Dennett 1994), and as Lycan doesn't criticize my version of thin materialism, I don't find this criticism ample, but I must admit that Graham and Horgan are only going along with the mainstream in ignoring my brand of thin materialism.

That's why the current essay is necessary.

Graham and Horgan spell out the best of the thick materialist campaigns—Michael Tye's PANIC—and imagine their own variation on the original theme: Mary Mary, the daughter of the original Mary, and a devotee of Tye's brand of thick materialism.

According to Tye's PANIC theory, “phenomenal character is one and the same as Poised Abstract Nonconceptual Intentional Content” (Tye, 1995: 137), which means roughly that it is content that is “in position to make a direct impact on the belief/desire system” and is about nonconcrete, nonconceptualized discriminable properties. It follows from Tye's view, they claim, that Mary Mary, upon release, *shouldn't be surprised*. As they say: “In the end, Tye's version of thick materialism is just *too thin*. And this problem threatens to arise for any materialist treatment of phenomenal content” (2000: 77; emphasis in original).

I had previously viewed Tye's alternative to my brand of thin materialism as giving too much ground to the qualophiles, the lovers of phenomenal content, but thanks now to Graham and Horgan I can welcome him into my underpopulated fold as a thin materialist *malgré lui*, someone who has articulated much more painstakingly than I had just what sorts of functionalistically explicable complexities go to *constitute* the what-it-is-likeness, the so-called *phenomenality*, of conscious experience. I applaud Graham and Horgan's analysis of Mary Mary's predicament, which leads inexorably to the conclusion that since she already knows all the facts, has all the information needed to have anticipated *all* the noticeable, articulable properties of her debut experience in a colored world, she should not, in spite of Tye's claims, be (or expect to be) surprised. Here, in a nutshell, is what they say:

First, what is psychologically significant about the PANIC properties is just the functional/representational role they play in human cognitive economy—something that Mary thoroughly understands already, by virtue of her scientific omniscience. ...

Second, what is psychologically significant about phenomenal concepts (given Tye's theory) is that they are *capacity-based* concepts; ... But she already understands these

capacities thoroughly, including how PANIC states subserve them, even though she does not possess the capacities herself. No expected surprises there, either.

Third, the psychological distinctiveness of beliefs and knowledge-states employing phenomenal concepts is completely parasitic (given Tye's theory) upon the capacity-based nature of the phenomenal concepts. So she already understands well the *nature* of these beliefs and knowledge-states. ... So Mary Mary, as a True Believer in Tye's PANIC theory of phenomenal consciousness, has no good reason to expect surprise or unanticipated delight upon being released from her monochrome situation. (2000: 71–72) In short, Tye should join me in predicting that Mary Mary, like her mother Mary, would *not* be surprised or delighted at all. She's been there, done that, in her vast imagination already, and has nothing left to learn. So what's the problem? Why don't Graham and Horgan join Tye and me? (I'm presuming for the fun of it that Tye is now on my side.) Because—and here comes the super-pure, double-distilled intuition that I've been gunning for—“Surely, we submit, she should be both surprised and delighted” (2000: 72). “Surely.” As I noted in “Get Real” (Dennett 1994) in one of my many commentaries on Ned Block, “Wherever Block says ‘Surely,’ look for what we might call a mental block” (549). Block is perhaps one of the most profligate abusers of the “surely” operator among philosophers, but others routinely rely on it, and every time they do a little alarm bell should ring. Here is where the unintended sleight-of-hand happens, whisking the false premise
end p.18

by the censors with a nudge and a wink. Graham and Horgan do pause momentarily to ask why they are so sure, and this is what they answer:

What will surprise and delight Mary Mary ... is (it seems to us) the unanticipated *experiential basis* of her concept-wielding, recognitional/discriminatory, capacities and the acknowledged richness of her experience; she never expected polychromatic experience to be like *this*. (72)

I know that it seems to many people that there is this extra “richness,” this “*experiential basis*” over and above all the PANIC details, but I have claimed that they are just wrong about this, and I have offered a diagnosis of the sources of this deep-seated theorists' illusion. In “Quining Qualia” (Dennett 1988), I discussed the example of the torn Jell-O box, half of which has shape property M, and the other half of which is the only *practical* M-detector: the shape may *defy description*, but it is not literally ineffable or unanalyzable; it is just extremely rich in information. It is a mistake to inflate practical indescribability into something metaphysically more portentous, and I have been urging people to abandon this brute hunch, tempting though it may be. But Graham and Horgan cannot bring themselves to abandon the intuition. More important, they cannot even bring themselves to acknowledge that their whole case thus comes down to simply announcing their continued allegiance to a claim that, whether it is true or false, has been declared false and hence could use some support. They offer no support for it, but they do keep coming back to it, again and again:

Although phenomenal states may indeed play a PANIC role in human psychological economy, their phenomenal character is not reducible to that role. It is something more, something surprising and delightful. (73)

Who says? This is just what I have denied, at length.

Its greater richness is what is surprising and delightful about it, and Tye's theory leaves this out. (73)

This “greater richness” is just what needs to be demonstrated, not assumed. After all, the point of the Mary example is supposed to be that although thanks to her perfect knowledge she can anticipate *much* of what it will be like to see colors, she cannot anticipate it *all*. Since some of us have claimed that there is no reason to deny that all the “greater richness” *is* accessible to Mary in advance, this bald assertion by Graham and Horgan is question-begging. It simply won't do to lean on the obvious fact that under normal circumstances, indeed under any circumstances except the wildly improbable extreme circumstances of this thought experiment, Mary would learn something. But she *will* experience surprise and unanticipated delight, upon release from her monochromatic environment—which presumably should lead her to repudiate the materialist theory she previously accepted. (74)

So they say. Now thin materialism may, in the end, be false, but you can't argue against it by just saying “Surely not!” I have claimed that the richness we appreciate, the richness that we rely on to anchor our acts of inner ostension and

end p.19

recognition is *composed of* and *explained by* the complex set of dispositional properties that Tye has called PANIC properties. But Graham and Horgan make the mistake of assuming that there is, in addition to all this, a layer of “direct acquaintance” with “phenomenal properties.” They say baldly:

There is also direct acquaintance with phenomenal character itself, acquaintance that provides the experiential basis for those recognitional/discriminatory capacities. (2000: 73)

And also:

She claims to be delighted. ... Auto-phenomenology suggests strongly, *very* strongly, that she is right about this: the intrinsic phenomenal character of color experience is distinct from, and provides the basis for, these recognitional/discriminatory capacities. (77)

This, according to me, is just the reverse of what's true. These capacities are themselves the basis for the (illusory) belief that one's experience has “intrinsic phenomenal character,” and we first-persons have no privileged access at all into the workings of these capacities. That, by the way, is why we shouldn't do auto-phenomenology. It leads us into temptation: the temptation to take our own first-person convictions not as data but as the undeniable truth.

So on his [Tye's] story, Mary's post-release heterophenomenological claims evidently must be viewed as rationally inappropriate, and thus as embodying some kind of error or illusion. *That* is the basic problem: the apparent failure to provide adequate theoretical accommodation for the manifest phenomenological facts. (77)

The basic problem, they say, is dealing with these “manifest” facts, but it's only a problem if, in fact, she will learn something. It is not a problem for my view (and Tye's, if he'll join the thin materialists): she won't learn anything and won't be surprised; there are no such manifest phenomenological facts. At this point, if you are like many of my students, you are beset with frank incredulity. *Of course* Mary learns something on

release! She *has to*! Oh? Then please give me an argument, based on premises we can all accept, that demonstrates this. But I have never seen such an argument even attempted. “It stands to reason!” people say, and then they decline to offer any reasons, thinking them somehow uncalled for. I call for them.

In response to the previous paragraph in an earlier draft, Bill Lycan has answered the call: Here's a way to see why some of us think Mary does learn something. What one knows when one knows w.i.l. [what it's like] to experience a blue sensation is ineffable; at least, it's very tough to put into (noncomparative) words. One resorts to the frustrated demonstrative: “It's like ... *this*.” The reason physically omniscient Mary doesn't know what it's like is that the ineffable and/or the ineliminably demonstrative can't be deduced, or even induced or abduced, from a body of impersonal scientific information. (personal communication)

I daresay that Lycan speaks for many who are sure that Mary learns something, so now we have an explicit rendering of a background presumption of ineffability and an illustration of the role it plays in the argument I call for. Now what about
end p.20

that argument? First of all, nobody could deny that these propositions ventured by Lycan are large theoretical claims, not minimal logical intuitions or the immediate, unvarnished judgments of experience. *What one knows when one knows what it's like to experience a blue sensation is ineffable*. I suppose the concept of ineffability being appealed to here would get elaborated along these lines:

It is not the case that there is a string of demonstrative-free sentences of natural language, of any length, that adequately expresses the knowledge of what it is like to experience a blue sensation.

One would like to see that proved. (I'm being ironic. Of all the things one might want to construct a formal theory of, *ineffability* is way down the list, but it might be worthwhile to consider the difficulty of any such undertaking.) Presumably one wants to contrast the ineffability of what it's like to experience a blue sensation with, say, the ready effability (if I may) of what it's like to experience a triangle.

Someone who has never seen or touched a triangle can presumably be told in a few well-chosen words just what to expect, and when they experience their first triangle, they should have no difficulty singling it out as such on the basis of the brief description they had been given. They will learn nothing. With blue and red it is otherwise—that, at any rate, is the folk wisdom relied upon by Jackson's example. (He wouldn't have gotten far with a thought experiment about Mary the geometer who was prevented from seeing or touching triangles.) But if what it is like to see triangles can be adequately conveyed in a few dozen words, and what it is like to see Paris by moonlight in May can be adequately conveyed in a few thousand words (an empirical estimate based on the variable success of actual attempts by novelists), are we really so sure that what it is like to see red or blue can't be conveyed to one who has never seen colors in a few million or billion words?

What is it about the experience of red, or blue, that makes this task impossible? (And don't just say, They're *ineffable*. We are enjoined by the extremity of the thought experiment to take this seriously.) Remember, Mary knows *everything* about color that can be learned by physical science, and she presumably has the attention span and powers

of comprehension required to handle 10 billion words on what it is like to see red as easily as she does twenty-five words or less on triangles. Lycan says, “At least it's very tough to put into (noncomparative) words,” but this is not a thought experiment about difficulty; it's a thought experiment about impossibility. The fact that people find it hard to imagine that any description of what it's like to see red could do the job is negligible support. Faced with such a formidable task, one does indeed fall back on what Lycan aptly calls the “frustrated demonstrative,” but it is a long way from the undeniable claim that it is *very tough* to think of ways of characterizing what it is like without resorting to such private demonstratives, to the grand claim that such private demonstratives are, strictly speaking, ineliminable. And only absolute ineliminability would carry any weight in an argument against the *possibility* of Mary inferring what it would be like for her to see red. So I stick to my guns. The standard presumption that Mary learns something, that Mary *could not* have figured out just what it would be like for her to see colors, is a bit of folk psychology with nothing but tradition—so far—in its favor.
end p.21

You Had to Be There

Another unargued intuition exploited by the Mary intuition pump comes in different varieties, all descended inauspiciously from Locke and Hume (think of Hume's missing shade of blue). This is the idea that the “phenomenality” or “intrinsic phenomenal character” or “greater richness”—whatever it is—cannot be constructed or derived out of lesser ingredients. Only actual experience (of color, for instance) can lead to the knowledge of what that experience is like. Put so boldly, its question-beggingness stands out like a sore thumb, or so I once thought, but apparently not, since versions of it still get articulated. Here are two, drawn from Tye and Lycan:

Now, in the case of knowing via phenomenal concepts, knowing what it is like to undergo a phenomenal state type P demands the capacity to represent the phenomenal content of P under those concepts. But one cannot possess a predicative phenomenal concept unless one has actually undergone token states to which it applies. (Tye 1995: 169) ³

As Nagel emphasizes, to know w.i.l., one must either have had the experience oneself, in the first person, from the inside, or been told w.i.l. by someone who has had it and is psychologically very similar to oneself. (Lycan 2003: 389) ⁴

The role of this presupposition is revealed in the many attempts in the literature to guarantee that Mary doesn't cheat, somehow smuggling the experience of color into her cell. What special care must be taken to prevent Mary from taking surreptitious sips from the well of color? The blockades erected by Jackson in his original telling have long been recognized as insufficient as they stand. Mary might, for instance, innocently rub her closed eyes one day and create some colored “phosphenes” (try it—I just got a nice, deep indigo one right in the middle of my visual field). Or she might use her vast knowledge to engage in some transcranial magnetic stimulation of her color-sensitive cortical regions, producing even gaudier effects for her to sort out. Should a sophisticated alarm system be installed in her brain, to cut short any dream “in color” that she might innocently wander

into by happenstance? Is it in fact possible for a person to dream in color if that person has never seen colors while awake? (Whaddya think? Some might be tempted to respond: “Naw. The colors have to *get in there* through open eyes in order to be available for later use in dreaming.” That’s the Lockean premise laid bare, and presumably nobody would be seduced by it in such a raw form today.) If Mary’s color vision system is still intact—a nontrivial empirical assumption, given what is known about the ready reassignment of unused cortical resources in other regards—then she already has “in there” everything she needs to experience color; it just hasn’t been stimulated. (That, at any rate, is the stipulation on which the thought experiment depends, however unrealistic it may be empirically.) A dream could trigger the requisite activity as readily, presumably, as any external stimulus to the open eyes. There are no doubt myriad ways of short-circuiting the standard causal pattern and producing color experience in the absence of external world color.

More ominously for the prospects of the thought experiment, there are no doubt myriad ways of adjusting the standard causal pattern to produce some state of the brain that is almost the same as the sort of state that underlies standard color experience, but that differs in ways that subvert the clarity of the scenario and what it is meant to prove. What started out as a crisp, clean, “intuitive” predicament is being pulled out of shape by the inconvenient complications of science. According to the original thought experiment, it is the subjective, internal *experience* of color, however produced, that is held to be a prerequisite for knowing what it is like to see red, but now that this thesis has lost its naïve anchoring in eyes-open-and-awake, it cannot so readily be distinguished from other states of mind that have many of the effects of experiences of color without clearly being experiences of color. To take the most obvious case, if you right now imagine you are seeing a red rose, do you *thereby* experience red? (Here is an argument, if a need for one is felt: imagining anything is having an experience, so imagining a red rose is having an experience as of a red rose, which is different from having an experience as of a yellow rose, and the difference must be that in the former case of imagining, you have an experience of red.) As plausible as this can be made to seem, if it is endorsed, triviality looms for the Mary argument: to know what it is like to experience red is to imagine what it is like and imagine it correctly; but to imagine experiencing red just *is* to experience red, so it follows trivially that you can’t know what it is like to experience red until you have experienced red.

We are told that Mary in her cell can’t imagine what it’s like to experience red, try as she might. But suppose she doesn’t accept this limitation and does try her best, cogitating for hours on end, and one day she tells us she just got lucky and succeeded. “Hey,” she says, “I was just day-dreaming, and I stumbled across what it’s like to see red, and, of course, once I noticed what I was doing I tested my imagination against everything I knew, and I confirmed that I had, indeed, imagined what it’s like to see red!” Doubting her, we test her by showing her a display of three differently colored circles, and she immediately identifies the red one as red.⁵ What would we conclude?

end p.23

A. Jackson was wrong; Mary *can* figure out what it’s like to see red in the absence of any experience of red; or

- A. Jackson was wrong; Mary *can* figure out what it's like to see red in the absence of any experience of red; or
- B. Mary didn't *figure out* what it is like to see red; she had to resort to (highly intelligent, theory-guided) exercises of *imagining* in order to come to know what it is like to see red. By *imagining* red, she was actually illustrating Jackson's point, not refuting it. As her example shows, you can't know what it's like before you've actually experienced what it's like.

An awkward moment: a simple variation on the tale that clearly refutes it or clearly vindicates it, depending on how you interpret what happened. If B is the only conclusion Jackson intended, then we philosophers have been wasting a lot of time and energy on what appears to be a relatively trivial definitional issue: nothing is going to be allowed to *count* as a state of knowing what it's like to see red without also counting as an experience of red.

Before looking more closely at this contretemps, let's consider one other variation, one I would have thought was the obvious variation for philosophers: Swamp Mary.⁶

Suppressing my gag reflex and my giggle reflex, here she is:

Swamp Mary: Just as standard Mary is about to be released from prison, still virginal regarding colors and aching to experience “the additional and extreme surprise, the unanticipated delight, or the utter amazement that lie in store for her” (Graham and Horgan 2000: 82), a bolt of lightning rearranges her brain, putting it by Cosmic Coincidence into exactly the brain state she was just about to go into *after* first seeing a red rose. (She is left otherwise unharmed of course; this is a thought experiment.) So when, a few seconds later, she is released, and sees for the first time a colored thing (that red rose), she says just what she would say on seeing her *second* or *nth* red rose. “Oh yeah, right, a red rose. Been there, done that.”

Let me try to ensure that the point of this variation is not lost. I am *not* discussing the case in which the bolt of lightning gives Swamp Mary a hallucinatory experience of a red rose. That is, of course, one more possibility, but it is not the possibility I am introducing. I am supposing instead that the bolt of lightning puts Swamp Mary's brain into the dispositional state, the competence state, that an experience of a red rose would have put her brain into had such an experience (hallucinatory or not) occurred. So, after her Cosmic Accident, Swamp Mary may *think* that she's seen a red rose, experienced red, been in a token brain state of the type that subserves experiences of red, but she hasn't. It's just as if she had. Maybe she wrongly remembers or seems to remember (just like Swampman) having seen a red rose, or maybe, in spite of her lacking any such episodic memories, her competences are otherwise all *as if* she had had such episodes in her past. (After all, you could forget your first color experiences and still have phenomenal concepts, couldn't you?) *Ex hypothesi*, she didn't have any such experiences, whatever she now thinks; any bogus memories of color were inserted illicitly in her memory box

end p.24

by the lightning bolt. Hey, [surely] it's *logically* possible. Swamp Mary is exactly like Mary, an atom-for-atom duplicate of Mary at every moment of her life except for a brief interlude of lightning that performs the accidental (but not supernatural) feat of doing in a

flash exactly what Mary's looking at the rose would do by more normal causal routes. It follows that those who think "that there are certain concepts that ... can only be possessed and deployed on the basis of having undergone the relevant conscious experiences oneself" (Graham and Horgan 2000, speaking of Tye 1995: 65) may be right as a matter of contingent fact, but it is logically possible for one to acquire this enviable ability by accidental means. (These words stick in my throat, but I'm playing the game as best I can.)

RoboMary

We now have two routes to Mary's post-release knowingness: the Approved Path of "undergoing the relevant conscious experiences oneself" and the logically possible Cosmic Accident Path. The second path is a throwaway, not worth discussing. What *is* worth discussing is a third route to this summit: not a pseudomiracle but an ascent by good hard work: Mary puts all her scientific knowledge of color to use and *figures out* exactly what it is like to see red (and green and blue) and hence is not the least bit surprised when she sees her first rose. This third path is hard to imagine, certainly, and as we have just seen, its difficulty is complicated by the threat of a retreat into circularity. It is high time to make the task easier, mounting a positive account that just might convince a few philosophers that they really can imagine it after all. I'm here to help. I will begin with a deliberately simple-minded version, for clarity, and gradually add the complications that the disbelievers insist on. In the spirit of cooperative reverse-engineering, I'm numbering the knobs on my intuition pump, and adding comments on how the knob settings agree or differ from other models of the basic intuition pump.

1. RoboMary is a standard Mark 19 robot, except that she was brought on line without color vision; her video cameras are black and white, but everything else in her hardware is equipped for color vision, which is standard in the Mark 19.

Hold everything. Before turning to the interesting bits, I must consider what many will view as a pressing objection:

Robots don't have color experiences! Robots don't have qualia. This scenario isn't remotely on the same topic as the story of Mary the color scientist.

I suspect that many will want to endorse this objection, but they really must restrain themselves, on pain of begging the question most blatantly. Contemporary materialism, at least in my version of it, cheerfully endorses the assertion that *we* are robots of a sort—made of robots made of robots. Thinking in terms of robots is a useful exercise, since it removes the excuse that we don't yet know enough about brains to say *just* what is going on that might be relevant, permitting a sort of woolly romanticism about the mysterious powers of brains to cloud our judgment. If materialism is true, it should be possible ("in principle!") to build a material

end p.25

thing—call it a robot brain—that does what a brain does, and hence instantiates the same theory of experience that we do. Those who rule out my scenario as irrelevant from the outset are not arguing for the falsity of materialism; they are assuming it, and just illustrating that assumption in their version of the Mary story. That might be interesting as social anthropology, but is unlikely to shed any light on the science of consciousness.⁷ Back to knob 1. Just like Mary, RoboMary's *internal* equipment is “normal” for color vision but she is being peripherally prevented—from birth—from getting the appropriate input. RoboMary's black-and-white cameras stand in nicely for the isolation of human Mary, and we can let her wander at will through the psychophysics and neuroscience journals reading with her black-and-white-camera eyes.

2. While waiting for a pair of color cameras to replace her black-and-white cameras, RoboMary learns everything she can about the color vision of Mark 19s. She even brings colored objects into her prison cell along with normally color-sighted Mark 19s and compares their responses—internal and external—to hers.

This was something that Mary could do, of course, only somewhat more tediously—she had to watch black-and-white TV while conducting all the experiments she needed to get that admirably complete compendium of physical information. This suggests a modest improvement that could be made in Jackson's original experiment, in which Mary's eyes are declared normal, and the entire color blockade has to be accomplished with prison walls, confiscation of mirrors, white gloves, and so on. As various commentators have observed, such a world would still be an ample source of chromatic input—shadows, and the like, not to mention the different shades of “white.” It would have been a lot cleaner for Jackson's original telling if he had stipulated that Mary had a pair of camcorders with black-and-white eyepieces strapped over her eyes, peering at the world all her life like somebody videotaping her vacation in Europe. (Or, slightly more science-fictionally, he might have imagined Mary not imprisoned but with “filters” implanted on her optic nerves, permitting only black-and-white signals through.)

3. She learns all about the million-shade color-coding system that all Mark 19s have. We don't know that human beings have the same color-coding system. Probably they don't, but this is just a complication we can leave out; if Mary knows *everything*, she knows all the variations of human color-coding, including her own.
 4. Using her vast knowledge, she writes some code that enables her to colorize the input from her black-and-white cameras (à la Ted Turner's cable network) according to voluminous data she gathers about what colors things in the world are, and how Mark 19s normally encode these. So now when she looks with her black-and-white cameras
- end p.26

at a ripe banana, she can first see it in black and white, as pale gray, and then imagine it as yellow (or any other color) by just engaging her colorizing prosthesis, which can swiftly look up the standard ripe-banana color-number-profile and digitally insert it in each frame in all the right pixels. After a while, she decides to leave the prosthesis turned on all the time, automatically imagining the colors of things as they come into focus in her black-and-white camera eyes.

Isn't this simply the robot version of phosphenes and transcranial magnetic stimulation— forbidden ways of getting color experience into RoboMary? Or is it rather a way of dramatizing the immense knowledge of color “physiology” that RoboMary, like Mary, enjoys? What is either of them allowed to do with their knowledge? Is this a cheat or isn't it?

Let's turn the knob both ways, and see what happens. In the first, and simplest, setting, we declare that just as Mary is entitled to use her imagination in any way she likes in her efforts to come up with an anticipation of what it's going to be like to see colors, RoboMary is entitled to use her imagination, and that is just what she is doing. After all, no hardware additions are involved: she is just considering, by stipulation, what it might be like to see color under various conditions. (We can suppose she goes to the trouble of considering dozens of variant colorization codings, so she has entertained many different hypotheses about what it is like to see red and other colors and has settled, defeasibly, on the one she thinks is best.)

5. She wonders if the ersatz coloring scheme she's installed in herself is high fidelity. So during her research and development phase, she checks the numbers in her registers (the registers that transiently store the information about the colors of the things in front of her cameras) with the numbers in the same registers of other Mark 19s looking at the same objects with their color-camera eyes, and makes adjustments when necessary, gradually building up a good version of normal Mark 19 color vision.

In the case of RoboMary, it is obvious what sorts of use she can make of her knowledge about color and color vision in Mark 19s. Much less obvious, of course, is what use human Mary could make of *her* knowledge. But that just shows how treacherous the original intuition pump is; it discourages us from even trying to imagine the task facing Mary if she wants to figure out what it is like to see red.

6. The big day arrives. When she finally gets her color cameras installed, and disables her colorizing software, and opens her eyes, she notices ... nothing. In fact, she has to check to make sure she has the color cameras installed. She has learned nothing. She already knew exactly what it would be like for her to see colors just the way other Mark 19s do.

Locked RoboMary

Too easy! Now let's turn the knob and consider the way RoboMary must proceed if she is prohibited from tampering with her color-experience registers. I don't know how Mary

could be crisply rendered incapable of using her knowledge to put her own brain into the relevant imaginative and experiential states, but I can easily describe the software that will prevent RoboMary from doing it. In order to prevent this sort of self-stimulation skullduggery (if that is what it is), we arrange for RoboMary's color-vision system—the array of registers that transiently hold the codes for each pixel in Mary's visual field, whether seen or imagined—to be restricted to gray-scale values. This is simple: we arrange to code the gray-scale values (white through many shades of gray to black) with numbers below a thousand, let's say, and simply filter out (by subtraction) any values for chromatic shades in the million-shade subjective spectrum of Mark 19s—and we put unbreakable security on this subroutine. Try as she might, RoboMary can't jigger her “brain” into any of the states of normal Mark 19 color vision *or imagination*. She has all her hard-won knowledge of that system of color vision, but she can't use it to adjust her own hardware so that it matches that of her conspecifics. Her color-representing hardware is disabled.

This doesn't faze her for a minute, however. Using a few terabytes of spare (undedicated) RAM, she builds a model of herself and from the outside, just as she would if she were building a model of some other being's color vision, she figures out just how she would react in every possible color situation.

I find that people have trouble imagining just how intimate and extensive this “third-person” knowledge would be, so let's indulge in a few illustrative details, to help furnish our imaginations. She obtains a ripe tomato and plunks it down in front of her black-and-white-cameras, obtaining some middling gray-scale values, which lead her into a variety of sequel states. She automatically does the usual “shape from shading” algorithm, obtaining normal convictions about the bulginess and so forth, and visually guided palpation gives her lots of convictions about its softness. She consults an encyclopedia about the normal color range of tomatoes, and she knows that these gray-scales in these lighting conditions are consistent with redness, but of course nothing comes to her directly about color, since she has black-and-white cameras, and moreover, she can't use her book learning to adjust these values, since her color system is locked. So, as advertised, she can't put herself directly into the *red-tomato-experiencing* state. She looks at the (gray-appearing) tomato and reacts however she does, resulting in, say, thousands of temporary settings of her cognitive machinery. Call that voluminous state of her total response to the locked gray-tomato-viewing *state A*. This is a state of her knowing what it is like for her to see a gray tomato. Then she compares state A with the state that her model of herself goes into. Her model isn't locked; it readily goes into the state that any normal Mark 19 would go into when seeing a red tomato. And since this is her model of herself, it then goes into *state B*, the state she would have gone into if her color system hadn't been locked. RoboMary notes all the differences between state A, the state she was thrown into by her locked color system, and state B, the state she would have been thrown into had her color system not been locked, and—being such a clever, indefatigable, and nearly omniscient being—makes all the necessary adjustments and *puts herself into state B*. State B is, by definition, *not* an illicit state of color experience; it is the state that such an illicit state of color experience normally causes (in a being just exactly like her). But now she can know just what it is like for her to see a red tomato, because she has managed to put herself into just such a dispositional state—this is, of course, the hard-work analog of the

end p.28

miraculous feat wrought by the Cosmic Accident of the lightning bolt in the case of Swamp Mary.

Her epistemic situation when she has completed this Vast (Dennett, 1995: 109) but not infinite labor is indistinguishable from her epistemic situation in the case in which we allowed her to colorize her actual input—and it had been conceded that in that epistemic situation she had known what it is like to see red, but the case was thrown out for cheating. So there are no surprises for her when her color system is unlocked and she's given color cameras. In fact, when she completes her model of herself, down to the very last detail, she can arrange for it to take over for her locked onboard color system, a spare color system she can use much as Dennett used his spare computer brain in “Where Am I?” (in Dennett 1978). Remember, RoboMary knows all the physical facts, and that's a lot.

Objection: RoboMary can't put herself into state B, the state her model is driven into by its unlocked color system, because that state involves the wielding of what Tye calls “phenomenal concepts,” and these are strictly parasitic on actual phenomenal experiences, which they quote or reproduce, in effect, when they are exploited in such demonstrative thoughts as “*that* is what red looks like.” Part of having the competence that comes (normally) from experience is being able later to use demonstratives with internal referents of this sort.

Oh really? Why can't RoboMary form demonstratives that allude to the relevant states of her model, instead of her own locked color system? And why wouldn't they be just as good? Because they wouldn't have that extra *je ne sais quoi*? But that is just what has not been shown to exist. In the case of RoboMary, the temptation to posit a rather magical extra property that adheres somehow to her entering into these color-system states (which are basically just numbers in registers, after all) is weak. The temptation should be resisted in the case of Mary, too. It has no legitimate business to do and tends to distort the imagination covertly.

Objection (thanks to the editors of this volume): For RoboMary to self-program herself into state B is cheating just as much as for her to self-program herself into the “experiencing red” state. What matters is whether Mary (or RoboMary) can *deduce* what it's like to see red from her complete physical knowledge, not whether one could use one's physical knowledge in some way or other to acquire knowledge of what it's like to see in color.

I just don't see that this is what matters. So far as I can see, this objection presupposes an improbable and extravagant distinction between (pure?) deduction and other varieties of knowledgeable self-enlightenment. I didn't describe RoboMary as “programming” herself; I said she “notes all the differences between state A, the state she was thrown into by her locked color system, and state B, the state she would have been thrown into had her color system not been locked, and—being such a clever, indefatigable, and nearly omniscient being—makes all the necessary adjustments and *puts herself into state B*.” If I use my knowledge to imagine myself into your epistemic shoes in some regard, is this “self-programming”? And if so, what is the import of this characterization for the knowledge argument? Consider Rosemary, another of Mary's daughters, who is entirely

normal and free to move around the colored world, and is otherwise her mother's equal in physical knowledge of color. Rosemary has a hard time imagining her mother's epistemic end p.29

predicament. What must it be like, she wonders, not yet to know what it is like to see red? She is burdened, it seems, with *too much* knowledge (cf. my example of the newly discovered Bach Cantata in *Consciousness Explained*, 1991: 388). This is, presumably, a psychological impediment to her imagination, but not an epistemological lack.

I take the example of RoboMary to shift the burden of proof. Thin materialism, the view that Mary, in her well nigh unimaginable circumstances, would not be surprised after all, has a lot to be said for it. Enough, surely, to undermine the blithe confidence with which philosophers have presumed otherwise.

A closing observation: I find that some philosophers think that my whole approach to qualia is not playing fair. I don't respect the standard rules of philosophical thought experiments. "But Dan, your view is so *counterintuitive*!" No kidding. That's the whole point. Of course it is counterintuitive. Nowhere is it written that the true materialist theory of consciousness should be blandly intuitive. I have all along insisted that it may be *very* counterintuitive. That's the trouble with "pure" philosophical method here. It has no resources for developing, or even taking seriously, counterintuitive theories, but since it is a very good bet that the true materialist theory of consciousness will be highly counterintuitive (like the Copernican theory—at least at first), this means that "pure" philosophy must just concede impotence and retreat into conservative conceptual anthropology until the advance of science puts it out of its misery. Philosophers have a choice: they can play games with folk concepts (ordinary language philosophy lives on, as a kind of aprioristic social anthropology) or they can take seriously the claim that some of these folk concepts are illusion generators. The way to take that prospect seriously is to *consider* theories that propose revisions to those concepts.

Acknowledgments

I am grateful to Diana Raffman, Bill Lycan, Victoria McGeer, and my students for many discussions, on e-mail and in person, on the ins and outs of this argument.

References

- Dennett, D. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge: MIT Press.
- Dennett, D. (1988). Quining Qualia. In *Consciousness in Contemporary Science*, ed. A. Marcel and E. Bisiach: 42–77. Oxford: Oxford University Press.
- Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown.
- Dennett, D. (1994). Get Real: Reply to My Critics. *Philosophical Topics* 22: 505–68.
- Dennett, D. (1995). *Darwin's Dangerous Idea*. New York: Simon and Schuster.

Dennett, D. (1996). Cow-Sharks, Magnets, and Swampman. *Mind and Language* 11: 76–77.

Graham, G., and Horgan, T. (2000). Mary Mary, Quite Contrary. *Philosophical Studies* 99: 59–87. [Link](#)

Hofstadter, D. (1981). Reflections. In *The Mind's I*, ed. D. Hofstadter and D. Dennett: 373–82. New York: Basic Books.

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36.

[Link](#)

end p.30

Lycan, W. (1996). *Consciousness and Experience*. Cambridge: MIT Press.

Lycan, W. (2003). Perspectival Representation and the Knowledge Argument. In *Consciousness: New Philosophical Essays*, ed. Q. Smith and A. Jokic: 384–95. Oxford: Oxford University Press.

McConnell, J. (1994). In Defense of the Knowledge Argument. *Philosophical Topics* 22: 157–87.

McGeer, V. (2003). The Trouble with Mary. *Pacific Philosophical Quarterly* 84: 384–93. [Link](#)

Robinson, H. (1993). Dennett on the Knowledge Argument. *Analysis* 53: 174–77.

[Link](#)

Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge: MIT Press.

end p.31

two So This Is What It's Like A Defense of the Ability Hypothesis

Laurence Nemirow

Many efforts have been made to debunk or refute the ability hypothesis (hereafter, AH), the theory that knowing what an experience is like may be identified with abilities to imagine, recognize, and remember, as opposed to propositional knowledge.¹ The eagerness to attack AH undoubtedly is driven by naked intuition. As David Chalmers (1996) has said, “Its *main* problem is that it is deeply implausible” (144; emphasis added). Such intuition-backed objections are worth only so much, however, for intuitive certainty, even when useful as a starting point, is an inherently unreliable guide to philosophical insight. “What we are tempted to say,” Wittgenstein (1958) pointed out, “is, of course, not philosophy—but its raw material” (§ 254).

Of more interest are the objective grounds for and against AH, which was developed as a response to the Knowledge Argument (hereafter, KA)—a purported proof of the existence of “phenomenal information” (as defined below). Given the reasoned nature of this debate, it betrays the spirit of the inquiry to base an attack on AH on the ground that it is intuitively unsatisfying (although I shall consider a variation of AH that is designed to avoid this supposed failing).

Phenomenal versus Physical Information

Central to the dispute between AH and KA is the concept of “phenomenal information,” which signifies information about experience that is not “physical information” in the broadest sense of that term.² “Physical information” encompasses (roughly) information expressible in the languages of physics, chemistry, and biology.

The knowledge argument purports to demonstrate the existence of phenomenal information that is learned by having experience. To do this, it posits black-and-white Mary, a scientist who knows everything science can ever teach about color vision (i.e., all relevant physical information), but not what seeing color “is like” because she has always lived in a colorless world. Despite her mastery of the science of color vision, KA asserts, Mary learns a new fact when she is first visually exposed to the color red (namely, what it's like to see red) that cannot be physical information (given that Mary already knew all the science).³ By contrast, AH grants that Mary learns something new, but maintains that her new knowledge is nonpropositional know-how rather than anything factual.

Underlying AH is the observation that although KA professes to display knowledge of what it's like to see red that is not expressible in physical terms, KA merely assumes that such knowledge is on a par with knowledge of physical information—that is, that the former, like the latter, is propositional knowledge. AH counters KA by adducing reasons against the soundness of this assumption and by arguing that knowledge of what it's like to see red is nothing but practical knowledge.

Overview of the Lewis Abilities

In his account of AH, David Lewis (1983, 1988) famously identified knowing what an experience is like with the practical knowledge that consists of the interrelated abilities to imagine, remember, and recognize the experience. I will describe these “Lewis abilities” in the context of color experiences.

For purposes of AH, the ability to imagine a color experience amounts to the ability to see a color in the mind's eye—to visualize it. People normally acquire this imaginative ability from firsthand (or personal) memories of color experience. A gardener tending a plot, for example, may cultivate a particularly striking rose, and, by seeing it for the first time, may develop the ability to remember and to visualize its particular shade of red.

For purposes of AH, the ability to remember can be analyzed as two abilities: the ability to visually remember having a color experience (to remember an experience of seeing a red rose) and the ability to visually remember the experience of visualizing a color (to remember the experience of visualizing rose red).⁴ The exercise of each of these abilities requires an act of visual imagination, although the reverse is not always true. All such remembering is imagining, but not all such imagining is remembering. Remembering may be regarded as a special context for imagining.

Finally, the relevant ability to recognize a color experience is the ability to distinguish it from others—to single out particular experiences in a continuum of color experiences that are connected through the relationship of similarity. A gardener

end p.33

exercises this ability in the course of comparing the red of the rose that she is (currently) tending to the color of the (imagined) rose that she aspires to grow or to the color of the (remembered) prize rose that grew in her mother's garden. As this example shows, imaginative and mnemonic abilities—such as the ability to visualize rose red and the ability to remember seeing a certain red rose—can also function to distinguish and identify experiences.

Objections to AH from Gifted Visualizers and Impaired Observers

A threshold contention of AH is that knowing what it's like to have an experience is coextensive with possession of the Lewis abilities. Rejecting this contention, Earl Conee (1994) deploys two thought experiments to show that the ability to imagine having an experience of a certain kind is neither necessary nor sufficient for knowing what it's like to have that kind of experience.

Against the sufficiency claim, Conee posits Martha, a woman who is highly skilled at visualizing intermediate shades of color that she has never seen by interpolating between colors she can remember having seen. Informed that cherry red is a shade of red midway between burgundy red and fire engine red, Martha is able to imagine cherry red.

However, at the moment before she first visualizes that shade, “it is clear that Martha does not yet know what it is like to see something cherry red” (Conee 1994: 138). She cannot gain such knowledge, according to Conee, before she actually imagines cherry red.

To argue that imaginative ability is not necessary for knowledge of what it's like, Conee posits a woman, call her Betty, who lacks all visual imagination but who nonetheless “clearly knows what it is like to see something red” while intently staring at a red tomato.⁵

Martha poses no problem in principle for AH.⁶ To deal with Martha, we can agree with Lewis that knowing what an experience is like depends not only on the ability to visualize the experience but also on the ability to remember—itself an imaginative ability. More precisely, Martha has the ability to know what it is like to see cherry red after she is able to remember visualizing cherry red. The latter mnemonic ability confers upon Martha knowledge of what it is like to see cherry red after she first visualizes it, even though she has never seen that shade of red and so cannot remember seeing cherry red (Conee 1994, n. 8).

As to Betty, Conee fails to address a risk inherent in his strategy of hypothetically severing important components of an individual's mental apparatus and making assumptions about what remains. The risk is that such radical hypothetical alterations will have major unintended consequences.

By stripping Betty of all ability to imagine color, Conee may have inadvertently denied her the knowledge at issue—namely, knowledge of what it's like to see red

end p.34

while intently staring at a red tomato. When we attribute knowledge of what it's like to see tomato red to ordinary people who are staring at a red tomato, we assume that they can activate a panoply of imaginative abilities. For example, seeing a red tomato, I can compare its hue to other colors that I can visualize or remember; I can imagine that the redness of the tomato occupies a larger or smaller portion of my visual field than it actually does; and I can imagine variations on the tomato's redness. If I were unable to activate any such abilities, you should be reluctant to agree that I know what it's like to see tomato red. So incapacitated, I would lack conscious awareness of the tomato's color. More generally, knowing what an ongoing experience is like by virtue of having that experience entails, if nothing else, conscious awareness of the experience, which itself involves abilities to *reflect* upon the experience, not merely to endure it—or in the particular case of the tomato, to stare intently at its redness.⁷

Moreover, deprived of the ability to imagine colors, Betty must also lack visual memory of colors, since visual memory implicates visual imaginative abilities. Lacking visual memory, Betty cannot visually remember seeing the color red from moment to moment—even as she sees it. So restricted, Betty's experience lacks the coherence that characterizes conscious awareness, and she cannot be said to know what it is like to see red for this reason, too.

In summary, AH asserts that imaginative abilities are essential to knowledge of what it's like to see red. Conee begs the question against AH by assuming that Betty, deprived of these abilities, “clearly” knows what it is like to see something red while she intently stares at something red.

What would someone be like who really knows what it's like to see a color only while seeing it? Consider Rex, who has normal abilities to recognize colors visually, and who can remember seeing colors from moment to moment so long as he is looking at samples of them, but whose visual memories of *any* color and related imaginative abilities fade away almost as soon as he stops looking. (Perhaps seeing a color temporarily jogs his memory of that color.) Of Rex we can correctly say that he knows what it is like to see colors only when looking at color samples (and for a few moments more), but not after his memory fades, even though at all times (including after his memory fades) he could recognize colors if color samples were placed within his field of vision.

Objection from Knowing with Particularity in the Moment

Michael Tye (2000) raises objections similar to Conee's to the AH correlation between knowing what it's like and the Lewis abilities. While looking at a red rose for the first time, Mary knows what it is like not merely to see red, but to see the particular hue of red before her (call it “red₁₇”). But she does not (Tye claims) obtain the related Lewis abilities because she cannot without further schooling

end p.35

reliably recognize, imagine, or remember that particular shade of red. Therefore, according to Tye, although Mary knows what seeing red₁₇ is like while looking at the rose, she has not acquired the Lewis abilities in question, and AH must be false.

Tye is mistaken, however, in asserting that his Mary lacks the relevant Lewis abilities while looking at a sample of red₁₇. Consider what abilities Tye's Mary must have in order to know what it is like to see red₁₇ while seeing red₁₇. First, she must be able to reliably distinguish red₁₇ from samples of red₁₆ and red₁₈ that are placed before her next to the sample of red₁₇. If she lacked this ability, we would be inclined to deny that she knows what red₁₇ looks like even while looking at red₁₇, since she can't recognize it. (We might agree, however, that she knows what it's like to see a more broadly defined spectrum of color that could be designated "red₍₁₆₋₁₈₎".)

Moreover, like Betty's knowledge of tomato red, Mary's knowledge of what it's like to see red₁₇ while staring at a sample of red₁₇ depends critically upon her present ability to imaginatively manipulate the color (her present ability to imaginatively vary the hue, and so on). In addition, while seeing red₁₇, Mary must be able to remember (at that moment and from moment to moment) seeing red₁₇ in order to have the conscious awareness of experience that is required to know what it is like to see red₁₇, although this ability may quickly lapse when she closes her eyes.⁸ In short, Tye's Mary must possess, for the moment at least, the requisite Lewis abilities to recognize, imagine, and remember red₁₇, or she would not know what it is like to see red₁₇ even while looking at a sample of red₁₇.

To further test AH's ability to handle Mary's experience with red₁₇, assume that some event diverts Mary's attention from the rose within her visual field, so that she does not (in this distracted state) know what it's like to see red₁₇.⁹ Distracted, Mary lacks the above-referenced recognitional abilities because she cannot recognize another patch of red₁₇ as red₁₇, and she cannot distinguish red₁₇ from red₁₆ and red₁₈. When distracted, she also can neither imaginatively manipulate her vision of red₁₇ nor remember how to imagine seeing red₁₇. Distraction interrupts Mary's recognitional, imaginative and mnemonic abilities, and this explains why she does not know what it is like to see red₁₇ while distracted.

In Tye's examples, Mary's knowledge of what it's like to see red₁₇ while seeing it is transitory. So too are the corresponding Lewis abilities because they survive only so long as Mary is exposed visually—and without distraction—to red₁₇. Thus, AH correctly predicts that Tye's Mary knows what it's like to see red₁₇ only while she sees it and only while she remains undistracted.

end p.36

Some Lessons of the Unusual Visualizers

Examples of unusual visualizers have suggested that Lewis was right to identify knowing what an experience is like with the contemporaneous possession of three related abilities—the abilities to imagine, remember, *and* recognize an experience. As Betty and Tye's Mary illustrate, two of the Lewis abilities—the abilities to imagine and remember—are necessary for what we think of as conscious awareness of experience, for, as I have argued, such awareness vanishes when those abilities are disabled. Though the Lewis abilities normally accompany one another, knowing what it is like begins to fail when they are separated from each other, as illustrated by Martha (who can

imagine cherry but cannot remember that color before she has either seen it or visualized it) and by Rex (who can recognize but can neither remember nor imagine color experiences while he is not seeing color).

The cases of Betty and black-and-white Mary also hold a lesson for the role that the Lewis abilities play in the language of experience, for Betty's and Mary's mnemonic and imaginative failures would prevent them from fully participating in ordinary conversation regarding color experience. People like Betty and black-and-white Mary in her color-deprived state are not able to associate color terms with their firsthand color experiences, and so cannot use these terms with what we think of as their full meanings. Though the Bettys and black-and-white Marys of the world might discuss the color experiences of other people, they lack the firsthand understanding of color terms that inform the meaning of those terms for ordinary people. (By comparison, Rex could have at least an episodic firsthand understanding of a color term, for he could associate the term "sky blue" with his own visual experience while looking at the sky on a clear day.)

As Betty and Mary illustrate, the Lewis abilities make available to language users firsthand meanings of color terms by allowing speakers to associate terms for experiences with their personal experiences. In particular, the ability to recognize colors visually, together with the mnemonic and imaginative abilities that similarly function to distinguish and identify experiences, enable language users to reliably associate terms for color experiences (such as "seeing rose red") with their own visualizations (e.g., firsthand visual memories of rose red). Conversely, an impairment of these Lewis abilities would represent an inability to draw upon the firsthand meanings of color experiences. (I will have more to say later about limitations on the meaning of color terms for imagination-impaired speakers.)

Objection from the Ability to Draw Inferences

Janet Levin (1990) finds it difficult for AH to explain why events of imagining can ground factual assertions about the world. As Levin says, "By being shown an unfamiliar color, I acquire information about its similarities and compatibilities with other colors, and its effects on other of our mental states; surely I seem to be acquiring certain facts about that color and the visual experience of it" (479).

It should come as no surprise, however, that abilities foster propositional knowledge. This is a garden-variety state of affairs. If I have expertise in dancing the waltz, I can predict the steps (or possible steps) within the dance from the beginning of a dance sequence; this does not mean that my dancing expertise is other than know-how. Similarly, with expertise in speaking English, I can evaluate whether purported sentences are well formed in that language; but this does not demonstrate that my English-speaking ability must be interpreted as propositional-based information. Inferences fostered by abilities are not really inferences at all, in the sense of following from or being justified by premises, so much as they are the product of reflexive tendencies (or abilities) to generate certain correct conclusions. In dancing the waltz as an

expert in the waltz, I just *know* the next move; I don't reason to reach it. If asked to explain how it is that I judge a weird English sentence (such as Chomsky's sentence, "fish who fish fish fish fish") as grammatical, I would be at a loss for words. I just *feel* that the sentence is grammatical; I don't infer that it is from any premises that I can state.

Challenged to justify the grammaticalness of the sentence, I would cite my own expertise as an English-speaker rather than articulate a theory of syntax. In the same vein, if I am shown a new shade of red, I can place it in a network of similar colors (e.g., between fire engine red and burgundy, to borrow Conee's example), but I do so almost instinctively, not by invoking any theoretical principles. If asked to justify a statement comparing the new shade to others, I might cite my experience with seeing color, or just reiterate my statement, for example, "because this shade of red *is* darker than that one and lighter than this one." (Black-and-white Mary could also compare colors in her color-deprived state, but her comparisons would be based on conclusions drawn from a complex assortment of physical information.) In general, the hallmark of reasoning from ability is the absence of the need to invoke a theoretical framework in order to validate conclusions.

Objection from Embedded Conditionals

Brian Loar (1990/97) objects to the technique of analyzing statements concerning experiences as abbreviated descriptions of know-how because this technique cannot explain how references to what experiences are like can be embedded in conditionals, such as: "If coconuts did not have *this* taste, then Q." There is no way, Loar asserts, to account for the embedded occurrence of "coconuts have this taste" as a matter of know-how, although he suggests that "you may get away with saying" that the statement "Coconuts have this taste" expresses "the mere possession of recognitional know-how." Loar is wide of the mark in criticizing AH for failing to analyze terms for experience (e.g., "Coconuts have this taste") outside of the scope of any knowledge qualifier. This criticism condemns AH for failing to satisfy a commitment that it never undertook—the commitment to analyze naked terms for experience.¹⁰ AH is
end p.38

designed to deal only with certain types of knowledge claims by reducing them to assertions about practical abilities. It asserts that statements about *knowing* what an experience is like are statements about abilities rather than claims to propositional knowledge. In short, AH is a theory about a certain kind of knowing, and nothing more. Defenders of AH may freely acknowledge that certain terms (such as "this taste") refer to experiences. No part of AH compels its defenders to deny that there are experiences; that we have a vocabulary to refer to them; or that they may be referenced by demonstratives.¹¹ Moreover, as a theory about a certain type of knowing, AH is not committed to the view that our naked references to experiences are reducible to statements about abilities. (Still to be considered, however, is the import for KA and AH of these acknowledged limitations on the ambitions of AH.)

Loar could reasonably challenge AH to address conditional statements with respect to *knowing* what it's like, such as: "If I didn't know that *this* is what it's like to taste coconuts, then Q." But a proponent of AH could treat "knowing that *this* is what it's like to taste coconuts" as equivalent to "being able to recognize, remember, and imagine *this* experience as the taste of coconuts," and could paraphrase the conditional statement as follows: "If I were unable to recognize, remember, and imagine *this* as the taste of coconuts, then Q."

Objection from Phenomenal Concepts

Martine Nida-Rümelin (1995) challenges proponents of AH to analyze statements that place phenomenal concepts in idiosyncratic epistemic contexts. In particular, she believes that the following statement poses difficulties for AH:

1. Marianna believes falsely that the sky appears red to normally sighted people. Marianna's story will help illuminate the nature of this challenge. Nida-Rümelin first supposes that Marianna, who is normally sighted (and believes that she is), lives (at all times through t_1) in a black-and-white environment. She is much like black-and-white Mary, but hasn't necessarily acquired Mary's scientific expertise. Later (at t_2), Marianna becomes acquainted with all the basic colors by viewing "artificial" objects (like walls and tables). At this stage, she is told the color names (like "orange" and "blue") that attach to many naturally occurring objects (like oranges or the sky, respectively), but she is not allowed to see these things, and she is unable to match a color with a color name. In fact, she somehow comes to believe that what people normally refer to as "blue" is the hue that we normally call "red." Thus, she knows that the term "blue" applies to the sky, but there is a clear sense in which she does not believe that the sky is blue. To identify this type
end p.39

of belief, Nida-Rümelin adopts a subscripting convention, under which the following is a true statement:

2. Marianna does not know at t_2 that the sky appears blue_p [i.e., blue phenomenally] to normally sighted people.

There is also a sense in which Marianna knows, after t_2 , that the sky appears blue to normally sighted people. Under her subscripting convention, Nida-Rümelin expresses this thought as follows:

3. Marianna knows at t_2 that the sky appears blue_{np} [i.e., blue nonphenomenally] to normally sighted people.

Finally (at t_3), Marianna is allowed to see the sky and acquires knowledge that the sky appears blue_p to normally sighted people.

Under the subscripting convention, Nida-Rümelin's challenge is more precisely stated as a request for a proponent of AH to analyze this statement:

4. At t_2 , Marianna believes falsely that the sky appears red_p to normally sighted people. Nida-Rümelin believes that AH cannot analyze (4) because Mary's false belief, as expressed in (4), results from the absence of factual knowledge, not from the absence of any ability.

A proponent of AH might consider paraphrasing (4) as follows:

5. At t_2 , if she were to attempt to visualize the color of the sky, Marianna would visualize red, falsely believing that she is visualizing the color that normally sighted people visualize when they visualize the color of the sky.

Alternatively:

6. At t_2 , of the activity of visualizing red, Marianna falsely believes that it is the same as visualizing the color that normally sighted people visualize when they visualize the color of the sky.

Moreover, at t_3 , the following become true:

7. In attempting to visualize the color of the sky, which in fact appears blue to normally sighted people, Marianna no longer mistakenly visualizes red.

8. When visualizing red, Marianna no longer falsely believes that she is visualizing the color of the sky, which in fact looks blue to normally sighted people.

Nida-Rümelin acknowledges the availability of a (roughly) similar type of analysis, but she denies that it helps AH because it “does not show that Marianna's progress [as of t_3] is not a genuine epistemic one that involves new knowledge of facts” (1995: 235–36).

Nida-Rümelin argues that Marianna acquires a new ability (i.e., to visualize blue in attempting to visualize the color of the sky) “only because she now has acquired the phenomenal knowledge needed for this practical ability” (1995: 235–36).

It is quite true, as Nida-Rümelin asserts, that Marianna, at t_3 , has learned new information, but this new information presents no problem for AH. One fact that she learns at t_3 may be expressed as follows:

end p.40

9. At t_2 , when Marianna attempted to visualize the color of the sky, Marianna in fact visualized red.

Alternatively:

10. At t_2 , when visualizing red, Marianna believed that she visualized the color of the sky.

However, the information expressed by (9) and (10) is not the kind of information that presents a problem for AH, namely, phenomenal information. To see this, note that black-and-white Mary (in her color-deprived state) could determine the truth of (9) and (10) through behavioral tests, perhaps by asking Marianna to pick out samples of the color that she imagines when she attempts to visualize the sky and comparing that color (by wavelength) to the color of the sky. Mary could also confirm (9) and (10) by undertaking a physiological inquiry.¹² Before t_3 —before personally observing the sky—Marianna herself might have learned the truth of (9) and (10) in a similar fashion or perhaps by being informed of their accuracy by a credible observer. This shows that the information reported by (9) and (10) is not the type of information posited by KA because its acquisition does not depend upon Mary's having had visual experience of colors.

What is indeed unavailable to both Mary and Marianna (in their color-deprived states) is the kind of knowledge that Marianna acquired at t_2 , and this may be described as knowledge of what it's like to see colors. To obtain this knowledge, visual exposure to colors was required; no number of words would have sufficed. Marianna's epistemic progress at t_3 , by comparison, did not require that she be exposed to the sky. That development could have been triggered by a verbal reference to what she already knew at t_2 —say, by a teacher telling her, “Marianna, the color of your kitchen wall is the same as the color of the sky!” This new knowledge does not support KA and is harmless for AH.

Objection from Mental Content

William Lycan (1996) contends that AH cannot do justice to this uncontroversial fact about imagining: “There is such a thing as getting ‘what it's like’ right, representing truly rather than falsely, from which it seems to follow that knowing ‘what it's like’ is knowing a truth” (99).

Is Lycan right to assume that AH is at odds with the assertion that in imagining we successfully represent the thing imagined? Suppose for the sake of argument that one indeed conjures up a mental *representation* of the state of seeing red when one visualizes red. On this assumption, a proponent of AH may explain that to

end p.41

know what it's like to see red, one must know how to imaginatively conjure up a state that *represents* the state of seeing red. Let's call this ability the “Conjuring Red Ability,” and let us call this variation of AH “Conjuring Red AH.” Of course, as a version of AH, Conjuring Red AH is intended to be consistent with (indeed, to be a special case of) AH. The original proponents of AH expressly made room for Conjuring Red AH.¹³ They

(we) never imagined that a representational theory of imagination is inconsistent with AH, and the existence of such an inconsistency should not be taken for granted. Conjuring Red AH performs a useful rhetorical function because it should capture the intuitive appeal of the idea behind phenomenal information: that what one does when one imagines red is to conjure up a mental representation of seeing red. If one shares this intuition, it should be equally intuitive to propose that knowing what seeing red is like is knowing *how* to conjure up such a mental representation.

Proponents of KA may assert that the Conjuring Red Ability itself provides access to phenomenal information. In conjuring up a mental picture of seeing red, I access a type of *content* that gives me access to the essential features of seeing red. As Lycan says, it is “what we succeed in imagining” that gives us information about the thing imagined.

“And, one would think, contents that afford inferences to propositional conclusions are themselves propositional” (1996: 96).

However, the assumption that representational content is propositional does not justify the conclusion that the content qualifies as “phenomenal information.” Tye explains this well:

From the fact that abilities with which knowing what it is like are identified are abilities to be in certain propositional states, it certainly does *not* follow that knowing what it is like is knowing a truth. What follows is that knowing what it's like consists in abilities, the exercise of which demands (at the time of exercise) the representation of certain truths. So what? (Tye 2000: 9)

It also does not follow that any of the information represented by any propositional state S that an imager represents (when he exercises the ability to visualize red) cannot be known to Mary (in her color-deprived state). We know that, whatever S may be, Mary cannot (in that state) imaginatively represent S. What has not been established, however, is that imaginative representation is the only way to know the information that S represents, or that Mary (in her color-deprived state) cannot know such information. The ability to model a fact may be one way, but not the only way, to come to know that fact.

¹⁴ Without making any commitment to elusive facts, Conjuring Red AH explains why having the ability to model (conjure up) mental representations of the color red is the only way to know what it's like to see red.

Knowing the Truth about Experience

But does black-and-white Mary really have the facts about color experience? I have made what might appear to be some concessions to the contrary. First, I have acknowledged (in the discussion of Loar) that propositions about one's own experiences and what one's own experiences are like genuinely refer to experiences. In addition, I have conceded that AH does not even attempt to reduce or explain away naked references to experiences, that is, references outside of knowledge qualifiers. Capitalizing on these acknowledgments, a proponent of KA might argue that knowledge of propositions about experience are what KA means by “knowing what it's like” and that this is the knowledge that Mary gains when she is released from captivity and allowed to see colors. ¹⁵ If correct, this argument shows that knowledge of the propositional expression of “what it's like” must elude Mary

in her captive state. Conversely, AH can survive this argument if it can be shown that black-and-white Mary indeed possesses such propositional knowledge or, if she does not, that all she is missing are the Lewis abilities.

Consider any proposition about experience the knowledge of which, according to a proponent of KA, is tantamount to knowing what the experience is like. Perhaps one such proposition is

(X) *This* is what it's like to see red
as uttered by Marianna when looking at her kitchen table. (If another candidate for (X) is preferred, (X) may be replaced in the argument that follows by any other proposition about experience knowledge that supposedly equates to knowledge of what it is like to see red. For example, (X) could be replaced in this argument by "*This* is the experience of seeing red" or "*This* is seeing red_p" in Nida-Rümelin's terms.)

To evaluate the import of proposition (X) for KA and AH, we must distinguish two senses of knowing the truth of (X).¹⁶ A person would know the truth of (X), in the "coarse-grained" sense, by having the right type of access to the right type of evidence to achieve knowledge of its truth without regard to the manner in which she is able to represent or apprehend that truth.¹⁷ Someone might achieve this type of knowledge without apprehending the truth of (X) from any specified perspective or under any specified description of the referent of "*This*."¹⁸

There are also "fine-grained" senses of knowing (X) to be true, although there is no unique or dominant "fine-grained" sense of such knowing. Someone would know (X) to be true in a "fine-grained" sense if she knew its truth from a particular
end p.43

perspective or under a particular description of the referent of "*This*." For example, only Marianna knows (X) from the perspective of the speaker of (X); and black-and-white Mary is among those who do not know (X) from the perspective of someone who has experienced the sight of red.

The knowledge that, according to KA, eludes Mary in her captivity must be construed as knowledge in the coarse-grained sense rather than in any fine-grained sense. The point of KA is not to assert that black-and-white Mary is ignorant of facts about what it's like to see colors only as understood from a specific point of view or only under specific representations. Rather, KA purports to falsify physicalism, which does not deny the truism that Mary gains new perspectives on experience when released from captivity.¹⁹ *Contra* physicalism, KA argues for the existence of facts that, regardless of the point of view from which they are apprehended or how they are represented, are simply out of Mary's purview. In short, what Mary learns on her release are supposed to be new facts *simpliciter*. (In Lewis's terms, KA asserts that Mary's new knowledge "eliminates possibilities" that her old knowledge did not preclude [1988: 507–09].)

Given KA's opposition to physicalism, a proponent of KA who endorses the view that knowledge of (X) is knowledge of what it's like to see red is committed to this proposition:

(Y) Before her release from captivity, black-and-white Mary cannot know the truth of (X) in the coarse-grained sense.

Yet it is difficult to see how (Y) can be true considering all the science that Mary knows during her captivity. Given her complete knowledge of physiology, Mary could determine when a person is normally sighted and what physical conditions are required for a normally sighted person to experience seeing red; and Mary could thereby come to know (at least in the coarse-grained sense) when Marianna is experiencing what it is like to see red, and so to know the truth of (X). Thus, after due inquiry into the physical states of Marianna's brain, Mary should be able to confirm (X) or, perhaps, to respond to (X) with a statement such as, "No, Marianna, *that* is not what it is like to see red, that is what it is like to see blue"—in either case without herself having the slightest idea of what it is like to see red or blue.

In response, it might be argued that Mary should not assume, based on physical facts alone, that what Marianna is experiencing is what others normally experience when they see red, even if Marianna is in the same physical state that others occupy when they see red. In other words, in reasoning about Marianna's experiences (and what those experiences are like), Mary may not assume that facts about experience supervene on the physical facts. This argument proves too much, however: it would never allow Mary to draw inferences about what Marianna's experiences are like, even after Mary is released from captivity. Even after Mary sees red, her only reason for believing that Marianna has experiences like her own are the physical reasons.

end p.44

Indeed, black-and-white Mary's own visual history should not constrain Mary's ability to ascertain what experiences Marianna is having at the moment. Assume, for example, that pre-release Mary observes Marianna experiencing the sight of red and that Mary confirms physically that this is in fact what Marianna is doing. Assume further that Mary is allowed to see red while she continues to observe Marianna experiencing the sight of red. At this juncture, Mary might well exclaim, "Now I know what it is like to see red!" or equivalently, "Now I know what you are experiencing!" She would not say, however, "Now I know that you are experiencing what it is like to see red"—either Mary already knew that, even before she herself knew what it is like to see red, or she did not know it and could not come to know it by experiencing the sight of red.

Taking a different tack, an opponent of AH might insist that someone like black-and-white Mary, who has not experienced the sight of red, simply is unprepared to understand what makes (X) true even if she could somehow confirm its truth.²⁰ Though Mary might know that Marianna tells the truth when Marianna says, "*This* is what it is like to see red," she does not and cannot understand what makes that assertion true. This failure of understanding is so significant that it arguably represents a true ignorance of a state of affairs—not merely a failure to comprehend a fact from a particular point of view or under a certain description.²¹

Though plausible at first blush, this last argument begs the question against AH. As we shall see, AH is compatible with an informative account of the type of understanding that black-and-white Mary does not have. (As elaborated in the next section, Mary's failure of understanding can itself be identified with the absence of the Lewis abilities.) If the AH-compatible account of such understanding is correct, then Mary's failure to understand

(X) represents no gap in her factual knowledge. So it would beg the question to assume that Mary's failure of understanding represents such a gap.

To turn the tables, a proponent of KA might charge AH with begging the question against KA by assuming that the Lewis abilities alone, without phenomenal information, provide the cognitive background required for the understanding of (X). According to this charge, knowledge of phenomenal information explains the acquisition of the Lewis abilities, and therefore the Lewis abilities may not be invoked to explain away the understanding that underlies such knowledge.²² However, this defense reveals a weakness in the very position that it tries to secure. Although KA starts out as an argument against physicalism, in its current
end p.45

incarnation it is reduced to relying upon knowledge of phenomenal information to “explain” the presence of the Lewis abilities.²³ If this “explanation” were known to be correct at the outset, there would be no need for KA in order to establish the existence of phenomenal information in the first place. Absent justification for such an “explanation,” Lewis was right, I think, to label it a “gratuitous metaphysical gloss” (1988: 517). In summary, we have found that black-and-white Mary could know the truth of (X) (in the relevant coarse-grained sense) before she knew what it's like to see red. Thus, knowledge of (X) (in that sense) could not constitute knowing what it's like to see red. What about the “fine-grained” senses in which black-and-white Mary may not know the truth of (X)? We can acknowledge that various perspectives on (X)—points of view from which Mary can know (X) to be true—elude her until after she is released from black-and-white captivity. For example, when released from captivity, Mary presumably comes to know that *she has experienced seeing red*; that *she can visualize red*; that *she can remember the experience of seeing red*; that *she can recognize an experience of seeing red*; and, in general, that she possesses the Lewis abilities relevant to knowing what it's like to see red. During her captivity, Mary could not know the truth of (X) from the point of view of someone who knows any of this information.²⁴ This is a consequence of AH—but not a problem for it—because the acquisition of such information is a natural result of Mary's acquisition of the Lewis abilities.

None of this suggests that any fine-grained way of knowing *that* is what constitutes knowing *what* an experience is like. To the contrary, the relevant fine-grained propositional knowledge arises from, and does not give rise to, the Lewis abilities. As noted above, the presence of the Lewis abilities creates the opportunity to acquire various fine-grained ways of knowing propositions about experience. Fine-grained ways of knowing such propositions in the absence of the Lewis abilities, however, do not lead to knowing what an experience is like. For example, having the point of view of someone who has seen red adds nothing useful to my knowledge that Marianna is currently seeing red—unless I can in fact remember the experience of seeing red (e.g., I have not forgotten how to remember). Similarly, having the point of view of someone who is currently seeing red would add no insightful perspective in the absence of the Lewis abilities: Unless I have visual imaginative and mnemonic abilities, my current experience of seeing red does not allow me to know what it's like to see red (as I have argued with respect to Conee's Betty and Tye's Mary).²⁵ The Lewis abilities, not the fine-grained ways of

knowing the truth about an experience, are what underlie knowing what an experience is like.
end p.46

Understanding Statements about Experience

Undoubtedly, something about Marianna's exclamation, "So *this* is what it's like to see red!" remains inaccessible to pre-release Mary even though she can test the accuracy of this exclamation. We want to say that Mary does not really understand the "content" of this exclamation, and she cannot do so precisely because she does not know what it's like to see red.

Proponents of AH are not without tools to address Mary's failure of understanding; for the same abilities that characterize knowledge of what it's like to see red also assist the understanding. In other words, the reason that pre-release Mary cannot fully grasp the meaning of third-person demonstrative attributions of color experiences (even though she may know full well when these attributions are accurate) is that understanding attributions of experience depends, to some extent, upon knowing how to imagine the experience.²⁶

Robert Gordon's theory of mental simulation generalizes this account of understanding attributions of experience. Gordon's idea, which has been subjected to abundant empirical testing (with results both favorable and unfavorable), is that the ability to engage in mental simulation is the key to understanding attributions of intentional states: "Only those who can simulate can understand an ascription of, e.g., belief—that *to S* it is the case that *p*."²⁷ Gordon's simulation theory, like AH, is an alternative to the hypothesis (the so-called theory theory) that propositional knowledge (knowledge "that") underlies human mental capacities. When the proposition *p*, in Gordon's formulation, is the proposition expressed by "This is what it is like to see red!" (as uttered by post-release Mary), Gordon's theory coincides with this special theory of understanding statements about what it's like: Only someone who can imagine seeing red can understand Mary's affirmation that "*This* is what it is like to see red." This special theory need not, however, draw upon Gordon's general theory for support. Gordon's theory is motivated by perceived limitations on the ability of the theory theory to account for how we predict, explain, and interpret the behavior of other people. By contrast, this special account of understanding propositions about experience derives its support from particular observations about how we use terms for experience.²⁸

A shared background assumption of language users is that (most) others can relate firsthand to experience by imaginatively thinking about named experiences. The shared assumption enables users of a language to tap into the (normally) shared network of imaginative abilities. For example, when we describe a hue of red, we (ordinarily) assume that others can remember, recognize, and imagine shades of color, and that we will evoke an imaginative response when we say something like: "Her skirt is a deeper shade of red than blood red." Someone who has never visually experienced the color red and therefore has not mastered the relevant imaginative abilities could not appreciate the image that statement is intended to evoke, even if she has mastered the science of vision

physiology. Similarly, someone who has never experienced pain could not fully understand reports of pain, even if she could confirm their truth based upon her mastery of pain physiology, because she could not imagine having a painful experience. As Wittgenstein observed, an inexperienced user of terms for experience would not be a full player in the language game of experience. He is like a child who says, "I don't know if what I have got is pain or something else" (1958: § 288). Witnessing this, we may think, "He does not know what the English word 'pain' means; and we should explain it to him ... [p]erhaps by means of gestures, or pricking him with a pin and saying: 'See, that's what pain is!' " (§ 288). We teach the child the meaning of the term "pain" by teaching him to associate that term with his own experiences of pain.²⁹

Know-How versus Ability

Noam Chomsky has drawn a distinction between know-how and ability, arguing that knowing how to speak and understand a language cannot be reduced to a system of abilities. His argument is based on a thought experiment. Imagine that Juan, a speaker of Spanish, suffers aphasia after receiving a severe head wound, losing all ability to speak and understand Spanish. Chomsky (1988) asserts that Juan has not lost his knowledge of Spanish, since he might recover his ability to speak and understand "as the effects of the injury recede." "Plainly, something was retained while the ability to speak and to understand was lost. What was retained was a system of knowledge, a *cognitive* system of the mind/brain" (10).

Chomsky has also applied this type of thought experiment to bike riding. He supposes that "Juan knows how to ride a bicycle, then suffers a brain injury that causes him to lose this ability completely ... then [he] recovers the ability as the effects of the injury recede. Again something remains unaffected by the injury that causes a temporary loss of ability. What remains intact was the cognitive system that constitutes knowing how to ride a bicycle; this is not simply a matter of ability" (11).

Borrowing and expanding on Chomsky's argument, Torin Alter (2001) distinguishes between knowing how to imagine having an experience and the ability to imagine the same experience. Suppose that a brain injury robs Hank of the ability to imagine the experience of tasting chocolate ice cream but that Hank's ability reappears as the injury recedes. Alter suggests that what is retained when the ability is temporarily absent must be knowledge of what it's like, which therefore cannot be reduced to abilities. "Hank's knowledge ... is implicit and temporarily inaccessible; nonetheless, *the knowledge is there*" (234, emphasis added).

The Chomsky/Alter approach to analyzing know-how implicitly assumes away the existence of deep-seated, structured abilities. What makes Chomsky and Alter so sure that Juan has completely lost the "ability" to speak Spanish just because he

end p.48

can't currently do so? Why not speculate instead that speaking Spanish is a complex, many-faceted, rule-bound ability, some aspects of which are relatively deeply embedded

in the brain of anyone who possesses the ability, and that Juan has lost some of this ability? What he has lost may be described roughly (and provocatively) as the ability to “exercise” his ability to speak Spanish;³⁰ or it may be described more generally as a peripheral component—some necessary but not sufficient element—of the ability to speak Spanish. One need not conclude, however, that Juan has entirely lost the ability to speak Spanish. Instead, one might conjecture that he has lost some, but not all, of what constitutes his Spanish-speaking ability, just as he has lost some, but not all, of his Spanish-speaking know-how. (After his accident, he might still know how to speak Spanish, but he certainly knows this less well than before.)

In the context of knowing what an experience is like, the Chomsky/Alter model of knowledge retention seems strangely out of touch. What makes KA so intuitively compelling is that it highlights the immediacy of our knowledge of color experience. When black-and-white Mary is first exposed to color, what she learns about what it's like to see red is vivid and obvious to her. If, shortly after she learns what it's like to see red, she suffers a brain injury that robs her of the ability to imagine red, what is left is not what we intuitively call “knowing what it's like to see red.” (We would find it odd, and perhaps a bit pathetic, if she were to continue to insist that she knows what red is like even though she has no better ability to imagine, recognize, or remember the sight of red than she had before being released from her black-and-white surroundings.)

Taking this point a step further, assume that black-and-white Mary is never exposed to color, but that a surgeon performs an operation that leaves her in the same brain-damaged condition that she would have been in after learning what it's like to see red and then temporarily losing the ability to imagine seeing red as a result of an accident. In other words, the surgery has the effect of endowing Mary with latent knowledge of what it's like to see red, which will become apparent to her as her brain recovers from the surgery. For some period of time after her surgery, she notices nothing strange or new; she has no immediate epiphany regarding what it's like to see red. Perhaps she does not even know that her brain has been altered. However, later in the recovery period, at time *t*, she finds that she can visualize red even though she's never had a red-seeing experience.³¹ We would, I think, all agree that, at time *t*, Mary finally learns what it's like to see red. However, Alter's position implies that after the surgery but prior to *t*—when her brain has been altered in some significant way, but the alteration has not yet afforded her the faintest idea of what it's like to see red—black-and-white Mary has made the transition to a person who knows what it's like to see red. However “implicit and temporarily inaccessible” the knowledge may be, “the knowledge is there” (Alter 2001: 234).
end p.49

Under Alter's analysis, knowing what it's like to see red has been transformed from one thing into something quite different. This type of knowing is known for its immediacy,³² but Alter characterizes it as potentially quite remote. Knowing what it's like to see red has become almost a theoretical entity—one that (at least in the unusual situations he constructs) we could know exists only because of its effects on observables (not because it is itself knowledge of an observable). Indeed, to highlight the potential inaccessibility of this knowledge, Alter labels it “X.” Alter tells us nothing about X other than (1) X is retained after an injury that robs Mary of her red-visualizing abilities; (2) X would seem

to have “the typical properties of knowledge”; and (3) the presence of X explains why Mary recovers the ability to imagine red experiences. So even though X represents the core component of knowing what it's like, Mary may fail to know that she possesses it. The nature of X is a mystery, but a different kind of mystery from what we have come to expect in discussing “what it's like to see red”; for X does not make itself manifest. This is a weird development—one that a proponent of phenomenal information should not welcome.

Conclusion

The character of the knowledge acquired by experience is deeply perplexing, yet most of us have powerful intuitions about what that knowledge is like. To cope with this tension, it may be helpful to recall these remarks from the *Tractatus Logico-Philosophicus*, although Wittgenstein undoubtedly wrote them with more spiritual matters in mind: “There are, indeed, things that cannot be put into words. *They make themselves manifest*. They are what is mystical” (1921: prop. 6.522; emphasis in original). While KA capitalizes on the mysticism inspired by “So *this* is what it's like”—and embraces the existence of propositional knowledge that “cannot be put into words”—AH provides a more worldly account that explains both the cognitive role of knowing what it's like and its essential connection with firsthand experience. Naturally enough, this approach engenders almost religious objection. But the test of a philosophical theory is not the fervor of the criticism it engenders, but the strength of the available rejoinders; and AH proves to be reasonably resilient to assault.

Acknowledgments

This chapter is dedicated to my father-in-law, Mortimer R. Kadish. I am indebted to comments provided by my wife, Joey Kadish, the editors, and the referee.

References

Alter, T. (1998). A Limited Defense of the Knowledge Argument. *Philosophical Studies* 90: 35–56.  [Link](#) ▶
end p.50

Alter, T. (2001). Know-How, Ability, and the Ability Hypothesis. *Theoria* 67: 229–39.
Anderson, S. (1988). *The Problem of the Unity of Consciousness*. Ph.D. diss., University of Colorado.
Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

- Chomsky, N. (1988). *Language and Problems of Knowledge: The Managua Lectures*. Cambridge: MIT Press.
- Conee, E. (1994). Phenomenal Knowledge. *Australasian Journal of Philosophy* 72: 136–50. [Link](#)
- Donnellan, K. (1977). The Contingent *A Priori* and Rigid Designators. *Midwest Studies in Philosophy* 2: 12–27.
- Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36. [Link](#)
- Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy* 83: 291–95. [Link](#)
- Gordon, R. (1986). Folk Psychology as Simulation. *Mind and Language* 1: 158–71. Reprinted in *Folk Psychology: The Theory of Mind Debate*, ed. M. Davies and T. Stone: 60–73. Oxford: Blackwell, 1995.
- Gordon, R. (2004). Folk Psychology as Mental Simulation. In *The Stanford Encyclopedia of Philosophy*, ed. E. Zalta. Available at: <http://plato.Stanford.edu/archives/fall2004/entries/folkpsych-simulation/>
- Kenny, A. (1984). *The Legacy of Wittgenstein*. Oxford: Blackwell.
- Levin, J. (1990). Could Love Be Like a Heatwave? Physicalism and the Subjective Character of Experience. In *Mind and Cognition*, ed. W. Lycan: 478–90. Oxford: Blackwell. First published in *Philosophical Studies* 49 (1986): 245–61.
- Lewis, D. (1983). Postscript to “Mad Pain and Martian Pain.” *Philosophical Papers*, Vol. 1: 130–32. Oxford: Oxford University Press, 1993.
- Lewis, D. (1988). What Experience Teaches. *Proceedings of the Russellian Society*. Sydney, Australia: University of Sydney. Reprinted in *Mind and Cognition*, ed. W. Lycan: 499–518. Oxford: Blackwell, 1990.
- Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives* 4: *Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview, 1990. Revised version in *The Nature of Consciousness*, ed. by N. Block, O. Flanagan, and G. Güzelde: 597–616. Cambridge: MIT Press, 1997.
- Lycan, W. (1996). *Consciousness and Experience*. Cambridge: MIT Press.
- Nagel, T. (1974). What Is It Like to Be a Bat? *Philosophical Review* 83: 435–50. [Link](#)
- Nemirow, L. (1980). Review of *Mortal Questions*, by Thomas Nagel. *Philosophical Review* 89: 473–77. [Link](#)
- Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. In *Mind and Cognition*, ed. W. Lycan: 490–99. Oxford: Blackwell.
- Nemirow, L. (1995). The Rules of Understanding. *Journal of Philosophy* 92: 28–52. [Link](#)
- Nida-Rümelin, M. (1995). What Mary Couldn't Know: Belief about Phenomenal States. In *Conscious Experience*, ed. T. Metzinger: 219–41. Exeter: Imprint Academic.
- Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press.
- Stich, S., & Nichols, S. (1992). Folk Psychology: Simulation or Tacit Theory? *Mind and Language* 7: 35–71.
- Tye, M. (2000). Knowing What It Is Like: The Ability Hypothesis and the Knowledge Argument. In *Reality and Humean Supervenience*, ed. G. Preyer and F. Siebert: 223–37. Rowman and Littlefield.

Wittgenstein, L. (1921). *Tractatus Logico-Philosophicus*. Translated by D. Pears and B. McGuinness. New York: Humanities Press, 1969.
Wittgenstein, L. (1953). *Philosophical Investigations*. Translated by G. E. M. Anscombe. London: MacMillan, 1958.
end p.51

three The Knowledge Argument, Diaphanousness, Representationalism

Frank Jackson

One good way of making a case against the knowledge argument is by noting that it conflicts with physicalism and rehearsing the very strong case for physicalism.¹ But this leaves unaddressed the undeniable force of the intuitions that drive the knowledge argument. I now think that the best strategy—the one that best enables us to see *where* the supporters of the knowledge argument, including my former self, went wrong—starts by isolating the key intuition that drives the knowledge argument and then showing how it conflicts with an attractive approach to phenomenal experience that can be independently motivated.

The Key Intuition behind the Knowledge Argument

I think the key intuition that drives the knowledge argument is that on leaving the black-and-white room, Mary acquires information about a new way that states are alike, one to another. When she leaves the room, she has certain highly distinctive experiences, and in consequence she acquires, it seems, an enlarged conception of the similarity patterns that obtain in our world. While in the black-and-white room, Mary knew that people are in various kinds of brain states that resemble each other in mass, chemical nature, temperature, functional roles, and so on. But, it seems, once she experiences red *as* red, once she knows what it is like to have that experience, she knows that there is something in common between states of subjects that outruns her previous knowledge; she learns a new way that certain items in our world—more particularly, certain experiences—resemble each other.

Many reply to the knowledge argument that what happens to Mary when she leaves the room is that she acquires new concepts, which is no reason to admit new properties; the knowledge argument fallaciously slides from the acquisition of new concepts to knowing about new properties.² I think the reason defenders of the knowledge argument find this reply unpersuasive is that the sense in which Mary would seem to acquire a new concept is that she learns of a new way of grouping experiences together. Someone who acquires the concept of, say, *charge* and learns that it applies to certain items learns of a new way that the items resemble each other, and that *is* to learn of a new property, the relevant unifier, that is instantiated in our world. In the same way, Mary's new concept seems to correspond to a new way for experiences to be alike, one that nowhere appears in the physicalists' picture; and if this is right, there are properties that fail to appear in that

picture, namely, those corresponding to her newly enlarged understanding of the respects of similarity that obtain between certain states of sentient creatures. My sense is that the example of water and H_2O has misled here. *Water* and *H_2O* are different concepts, and yet water *is* H_2O . This looks like good news for advocates of the view that “the knowledge argument confuses concepts and properties.” But Lavoisier *did* enlarge our understanding of what our world is like. The rise of modern chemistry told us new things about what kinds of properties are instantiated. This sets philosophers an interesting question. How should we give an account of the extra knowledge about the ways things are that came along with the discovery that water is H_2O while acknowledging the undoubted fact that water is H_2O ? But surely it would be wrong-headed to conclude that the rise of modern chemistry did not tell us new things about what our world is like. If I am right about the source of the intuitive force of the knowledge argument, the key contention that critics of the argument need to attack is the intuitively appealing one that Mary learns a new way in which certain items, in particular certain experiences, are alike. I think the best way to attack this contention—the “new similarity” contention, as I will sometimes call it—is via representationalism about sensory experience. More particularly, representationalism comes in different varieties, and it is the strong variety that undermines the key contention. In what follows, I first offer an argument for strong representationalism that takes off from the way diaphanousness shows a weaker version of representationalism to be untenable. This is the core of the chapter. I then spell out how strong representationalism undermines the knowledge argument—I think there has been a tendency to take this to be more obvious than it is—via the way it undermines the new similarity contention. I conclude by saying how I resist Torin Alter's argument in “Does Representationalism Undermine the Knowledge Argument?” (this volume, chap. 4) to the conclusion that representationalism does not undermine the knowledge argument. (I argue that he is right about one version of representationalism not undermining the knowledge argument but not about the strong version.)

end p.53

Representationalism: First Pass

We use predicates such as “square,” “red,” “in front of me,” and “stationary” to describe things in our world. We also use them to describe perceptual experience. We describe a table as square, in front of us, and brown. We describe our perceptual experience as being as of something brown, square, stationary, and in front of us. When psychologists in experiments ask us to describe how things seem, abstracting away from how we believe them to be, we use the same adjectives we use when saying how we believe things to be. It is obviously no accident that we give these words double duty. The question, What makes it right to use the word “square,” say, both to capture the nature of an object and to capture the nature of an experience? cries out for an answer. Representationalism explains this nonaccident by a certain kind of univocal thesis. To illustrate with the word “square”: it applies to something if and only if it has the property of being square; it applies to a visual experience if and only if the experience represents something as having the same property of being square. No special sense of “square” enters the story—

to be designated “square*,” as it might be when philosophical perspicuity is important—in order to account for why “square” applies to visual experience.

I am a convert to representationalism about perceptual experience (we won't be concerned with experience more generally, and “experience” unqualified in what follows should be read as the perceptual variety).³ And, as is the way with converts, I am eager to recruit. My efforts at recruitment in this chapter are, though, to some extent conditional and limited. They are limited to how the famous diaphanousness or transparency of experience can best be deployed to make an argument for representationalism, and they are conditional in that I largely assume diaphanousness.

Many have found diaphanousness very plausible (which is why I do not feel too bad about largely assuming it). Many have thought of it as the basis for a powerful argument for representationalism. I think, however, that the path from diaphanousness to representationalism has not been spelled out in the right way. Indeed, the usual view seems to be that diaphanousness, if accepted, is an argument in itself for representationalism. I start by explaining why I think that diaphanousness is in itself no argument for representationalism. As I argue in later sections, diaphanousness is, rather, an important intermediate premise (used twice over, as it happens) in the line of argument that takes us from what I will call weak representationalism to strong representationalism or representationalism proper—the kind of representationalism that, as I will later argue, shows us where the knowledge argument goes wrong by undermining the new similarity contention.

end p.54

Why Diaphanousness in Itself Fails as an Argument for Representationalism

That experience is diaphanous (or transparent) is a thesis about the phenomenology of perceptual experience.⁴ It is the thesis that the properties that make an experience the kind of experience it is are the properties of the object of experience. It is sometimes expressed by borrowing from Hume's famous remark about the self. Hume found himself unable to experience the self as such, always finding the experiences of the self getting in the way, so to speak.⁵ Likewise, it is plausible that we do not experience experience as such. The properties of the object putatively experienced always get in the way of attempts to access the phenomenology of experience itself. The claim is not, of course, that we cannot be aware that we are having such-and-such an experience or that there is no difference between the mental state of having such-and-such an experience and that of reflecting on that fact, and the like. The claim is that accessing the nature of the experience itself is nothing other than accessing the properties of its object.

Diaphanousness is very plausible, but our focus is on its implications for representationalism, and the trouble with using it as a launching pad for representationalism is that diaphanousness is *not* a claim about the nature of the object of experience *per se*. It is rather a way of affirming the famous act-object analysis that led so many to sense data. According to the act-object analysis, to have an experience is to stand in the relation of awareness to an object whose properties determine the kind of experience undergone. The contrast is with the adverbial analysis of sensory experience according to which to have an experience is to sense in a certain mode, where the mode determines the kind of experience undergone.⁶ But this means that diaphanousness says

nothing in itself that favors representationalism. One gets a consideration pointing toward representationalism only inasmuch as one has a reason to hold that the object of experience is an *intentional* object. If the object is an object in space-time, representationalism is false. In order for representationalism to be true, the object must be an intentional one—in particular, a way things are being represented to be. As we will see, there are good reasons to hold that the object is an intentional one, but this is no part of diaphanousness. It is an additional matter calling for separate argument. One way to see the point is to reflect on the fact that G. E. Moore (1903), perhaps the best known advocate of diaphanousness, used the argument as an argument for
end p.55

sense data, and sense data are *not* intentional objects. But the point is almost as obvious if you consider Gilbert Harman's presentation. He says:

When Eloise sees a tree before her, the colors she experiences are all experienced as features of the tree and its surroundings. None of them are experienced as intrinsic features of her experience. Nor does she experience any features of anything as intrinsic features of her experiences. And that is true of you too. ... Look at a tree and try to turn your attention to intrinsic features of your visual experience. I predict that you will find that the only features there to turn your attention to will be features of the tree. (1990: 667)

What, exactly, is the object that is claimed to have the “features” in this passage? If it is the tree, we do not have a generally acceptable account of what makes an experience the kind of experience it is. We know that the very same experience can be had in the absence of any physical object including trees.⁷ That is to say, the claim that “the only features there to turn your attention to will be features of the tree” is not in general correct. What is plausible in general is that whenever we try to “catch” the properties of our experience *qua* kind of experience that it is, all we seem to find are properties of *an* object whose nature determines the nature of the experience. But the nature of this object is a separate matter. Moreover, if one went by phenomenology—which is the basis for the claim that experience is diaphanous—the most plausible view straight off is that the object is an object in space-time and not an intentional object. We have learned to reject sense data, but there is a reason why they captivated so many for so long.

Or consider Michael Tye's recent discussion of diaphanousness—or transparency, as he calls it. Tye's view is that diaphanousness is “a very powerful motivation for the representationalist view ... but that the appeal to transparency has not been well understood” (2000: 45). He gives a detailed account in ten steps of how, in his view, we should spell out the path that takes us from diaphanousness to representationalism. Step 6 is the one of interest to us. After giving an account of what diaphanousness is and why it is plausible (a convincing account, as it seems to me), he says:

Step 6

What, then, is visual phenomenal character? One possible hypothesis is that it is a quality of the surface experienced. That hypothesis is intelligible only if it is assumed that the

surface is an immaterial one of the sort the sense-datum theorists posited. The best hypothesis, I suggest, is that visual phenomenal character is representational content of a certain sort—content into which certain external qualities enter. This explains why visual phenomenal character is not a quality of an experience to which we have direct access.
(48)

Instead of giving us answers as to how diaphanousness leads us to representationalism, it seems to me that this passage highlights the kinds of concerns we've
end p.56

raised. First, although it is widely and correctly assumed that the sense-datum theory is a mistake, to use its falsity as an unargued premise in an account of how diaphanousness leads to representationalism means that a key part of the account of why we should be representationalists does *not* rest on diaphanousness; it rests on the case against sense-data treated as a separate issue. Second, a lot of work is being done by the words “I suggest” in the quoted passage. It isn't clear here, or elsewhere in the ten-step argument as far as I can see, why diaphanousness per se warrants the suggestion “that visual phenomenal character is representational content of a certain sort”; but, in that case, diaphanousness is not doing the crucial work in the argument. Finally, the claim that “visual phenomenal character is not a quality of an experience to which we have direct access” seems false on reasonable understandings of the admittedly tricky notion of direct access, and, moreover, the claim is not something that follows from diaphanousness. Diaphanousness says that the properties of experience are the properties of the object of experience, not that we lack direct access to the properties.

I conclude that the famous diaphanousness or transparency of experience is not per se the basis of an argument for representationalism, even by the low standards converts are wont to set. It is not where we should start in developing the case for representationalism. We must look elsewhere for our starting point and, as I signaled earlier, bring diaphanousness into the argument along the way. The right place to start, in my view, is with the distinction between weaker and stronger versions of representationalism.

Weak, Minimal, and Strong Representationalism

Minimal representationalism holds that experience is *essentially* representational. Strong representationalism holds in addition that experience is *exhaustively* representational.

According to minimal representationalism, it is impossible to have a perceptual experience without thereby being in a state that represents that things are thus and so in the world, where “in the world” does not necessarily mean outside of the subject. Some experiences represent how one's stomach is, for example; but although this concerns how things inside one are, it concerns how the world is in the sense of concerning how things are with something distinct from the experience itself. Strong representationalism goes further in maintaining that how an experience represents things as being exhausts its experiential nature. It is not as if an experience's nature is partly constituted by how it represents things to be and partly by something else. How it represents things to be does the complete job. If experience consisted of a representational bit and a nonrepresentational extra, we could, strong representationalists argue, vary the “extra” while leaving the representational content unchanged. This would mean that we could

vary an experience's nature without varying how it represents things to be. And this, according to strong representationalism, is what cannot happen. Change an experience qua kind of experience it is, and you ipso facto change how it represents things to be.⁸ There is no extra element that might be tweaked in a way that leaves unaltered how things are being represented to be.

Minimal representationalism is consistent with and implied by strong representationalism. If how things are being represented is the sole determinant of experiential nature, experiences must by their very nature represent. It is useful to have a name for the kind of representationalism that affirms that experience is essentially representational while denying the exhaustion claim. I will call this view "weak representationalism." Thus, minimal representationalism comes in two forms: the strong version, which affirms exhaustion, and the weak, which denies exhaustion.

Two clarifications concerning strong representationalism. First, it is not the view that the *content* of an intentional state determines its nature qua mental state without remainder. That doctrine is false. A belief and a desire may have the very same content: I may both believe and desire that it will rain soon. Strong representationalism, as we will understand it, is the doctrine that the content of an experience *plus* the fact that the experience represents the content as obtaining in the way distinctive of perceptual representation are what determines the experience's nature without remainder.⁹ However, the difference between seeing red and seeing green is exhausted by content.

Second, it is important that it is *the* content, not part of the content, that appears in this formulation. A visual experience and a tactile one may equally represent that something is round, but they are very different experiences. Their difference lies, according to strong representationalism, in the fact that they have different contents; what they represent about how things are differs while agreeing in regard to the matter of shape. For example, the visual experience will represent how things are in regard to color while being silent about warmth and texture; the converse will be true of the tactile experience.

I take it that strong representationalism is the doctrine with bite: enough philosophers take it for granted that experience is essentially representational, that a perceptual experience by its very nature points to things being a certain way, for minimal representationalism to count as orthodoxy. Of course, how to analyze the relevant notion of representation is controversial. What I am saying is orthodoxy is the core idea that a perceptual experience by its very nature invites its subject to believe that things are a certain way.¹⁰ We may decline the invitation. We may indeed have no inclination to accept it.¹¹ All the same, if asked, Is there a way
end p.58

things are that one is being directed to as something the experience makes belief-worthy, absent defeaters?, it is very plausible that, necessarily for every experience, the answer is that there is such a way. After all, as we remarked at the beginning, the words we use to describe experience qua experience are the very words we use to describe the world and the things in it, and we take an experience we describe using those words to be in itself, albeit defeasibly, a reason for believing that how things are is that there are things in the world having the properties we use those words for.

I know there are dissenters to the kind of minimal representationalism I've called (tendentiously) orthodoxy, but I cannot think of an argument to persuade them to change their minds. Such an argument would need to have premises more plausible than that perceptual experiences are, by their very nature, representational, and that is a big ask in my book.

I now turn to the argument that takes us from minimal to strong representationalism or representationalism proper. I appreciate that the minority who dissent from minimal representationalism may take some comfort in the argument to come, seeing it as showing, as they see matters, that minimal representationalism is a wolf in sheep's clothing.

How Diaphanousness Takes Us from Minimal to Strong Representationalism

How might an experience essentially represent that things are thus and so? Any answer must advert to the nature of the experience. Something about the properties the experience has, in the sense of the kind of experience it is, makes it the case that it represents that things are thus and so. Let E be the relevant property of some experience in virtue of which it represents that the way things are has property P . We will review various possibilities for how E relates to P .

By diaphanousness, E is a property of the object of the experience. Is this object an object in space-time, presumably some kind of constituent of the experience, or is it an intentional object, presumably the very way that things might be, which is represented as being P ? Suppose the first. Then we have two sorts of problem. One sort is raised by the fact that in many cases E will have to be a property distinct from P . Sometimes our experience represents that something is square, and it is not plausible that the experience is, or has a part that is, square (except maybe by chance). The point is even more obvious for experiences that represent that something is a certain distance away. No part of the experience is some distance away from the subject.¹² Nor is it necessarily the case that the experience has a property that entails being square or some distance away. The properties E and P will typically be strongly distinct. But then how can it be that these distinct properties are *necessarily* connected, as must be the case if the experience's being

end p.59

E essentially represents that the way things are is P ? How can the instantiation of E *essentially* point to the instantiation of P ?

The second problem is independent of whether or not E and P are distinct properties. Suppose indeed that they are the very same property: $E = P$. How is it that an object in space-time's being P essentially represents that P is a property of how things are? We can understand how an object's being P might essentially represent that it itself is P , but the suggestion now under discussion is that an object's being P essentially represents that something *else* is P . As we noted earlier, minimal representationalism holds that experience essentially represents that the world is thus-and-so, where the reference to the world signifies that the representational content is directed to something other than the experience itself, and one thing's being thus-and-so does not in and of itself represent that something else is thus-and-so.

It can be tempting to think in terms of projection when we address the issue of how experience speaks to the nature of the world. The idea would be that when we have an

experience which is *E*, for suitable *E*, we project some property connected to *E*, the one we are calling *P*, which may or may not be *E* itself, onto the world. The experience represents that the world is *P* by virtue of the combination of being *E* and the act of projection. This, however, would not help with the problems just raised. First, is the act of projection part of what makes the experience the experience it is? If it is, we have a violation of diaphanousness. According to diaphanousness, it is the properties of the object of experience that settle the nature of experience, and projection is not a property of the object but instead is something done to certain properties of it. If, alternatively, the act of projection is not part of what makes the experience the experience it is, we have a violation of minimal representationalism. According to minimal representationalism, the experience's representing as it does is an essential part of its being the experience it is. It is not an extra consequent upon an act of projection conceived as distinct from what makes the experience the experience it is. Second, how can projecting properties from one thing to another be a matter of necessity, even if we have such qualifiers as that the projection be *prima facie* or *pro tanto* or ...? But in that case, a projection account is incompatible with the minimal representationalist's thesis that experiences, of necessity, point toward the world being thus and so.

The difficulties we have just surveyed arise from the assumption that *E*, the property that makes the experience the kind of experience it is, is a property of an object in space-time. In effect, we have used minimal representationalism plus diaphanousness in a *reductio* of any view that denies that the objects whose properties determine an experience's nature are intentional objects. Contraposing, we have shown that minimal representationalism plus diaphanousness implies that the properties of the experience are properties of an intentional object; they are properties of how things are being represented to be. The final step is to derive strong representationalism by a second appeal to diaphanousness. Diaphanousness says that the properties of the object of experience determine without remainder the nature of the experience. It follows that if the object of experience is an intentional object, the experience's properties are one and all the

end p.60

properties of how things are being represented to be. Here I mean the experience's properties *qua* kind of experience it is. As a good physicalist, I of course hold that the experience has all sorts of physical and functional properties that are not properties of an intentional object. Now talk of intentional objects should really have quotation marks around the word "object": the properties of an intentional "object" are nothing other than the properties of how things are being represented to be; they are, that is, properties of how things must be if things are to be as they are being represented to be.

We have, thus, reached strong representationalism, representationalism proper, the kind of representationalism that has the extra bite that weak representationalism lacks, by using diaphanousness twice over in an argument that presupposes minimal representationalism. The first use took us from minimal representationalism to the result that the objects that bear the properties are intentional objects. The second use delivered the exhaustion thesis distinctive of strong representationalism. We have reached the conclusion that the nature of a perceptual experience is exhausted by how it represents things to be from minimal representationalism plus diaphanousness.

I said earlier that diaphanousness is the wrong place from which to launch the case for strong representationalism. But of course my twofold use of diaphanousness to get from minimal representationalism to strong representationalism conforms with the thought that diaphanousness is crucial to seeing why we should be strong representationalists. I am dissenting from the letter of what many (strong) representationalists say while agreeing with a good part of the spirit.

How Strong Representationalism Undermines the Knowledge Argument's "New Similarity" Contention

Seeing red is a kind of experience, a highly distinctive kind. Attacks on qualia freakery and on the use of the phrase "what it is like" should not blind us to this evident fact. The intuition that fuels the knowledge argument—the new similarity contention, as we are calling it—is that Mary, in having that distinctive kind of experience, learns about a new kind of similarity holding between experiences. But what does that similarity consist in? If strong representationalism is true, there are two possible answers, for there are only two commonalities that might be relevant that obtain between different tokens of seeing red, to stick with that example, given strong representationalism. One is in how things are being represented to be; the other is in the fact that things are being so represented. The first commonality is in how things have to be for the experience to represent correctly; the second commonality is in the fact that each experience represents alike in the regard in question. I will argue that neither commonality makes trouble for physicalism. I will consider them in turn.

The challenge from the knowledge argument is the intuition that the "red" of seeing red is a new sort of property that unites the seeings of red. But commonalities in how things are being represented to be are not instances of properties. What unites how things have to be for the representations to be correct is not what
end p.61

unites the items that share the content. The "red" of seeing red cannot simultaneously be a property instance that Mary comes to know and what is shared by how things are being represented to be.

Here is a way to make the point via an argument that almost no one nowadays takes seriously. Suppose someone argued in the manner of the traditional argument from illusion against physicalism as follows.

1. When a straight stick immersed in water looks bent to degree d at some given time to me, its looking bent to degree d is to be understood in terms of its being bent to that degree.
2. Nothing physical is bent to degree d at that time, in front of me or in my head (we may suppose).
3. Therefore, there is at least one instance of something's being bent to a certain degree that physicalism fails to account for. Thus physicalism's inventory of which properties

1. When a straight stick immersed in water looks bent to degree d at some given time to me, its looking bent to degree d is to be understood in terms of its being bent to that degree.
are instantiated is incomplete.

Representationalism says that this argument goes wrong because the sense in which the first premise is true is one in which looking bent to degree d is understood in terms of there needing to be something bent to that degree for the visual experience to represent correctly, and not in the sense in which looking bent to degree d requires that being bent to that degree is anywhere instantiated. Mutatis mutandis for representationalism and the knowledge argument on the reading in which the commonality is in how things are being represented to be.

What about the alternative way of reading the similarity: as the similarity of being a state with a certain representational content? The contents are the same, but the “red” of seeing red lies, on this alternative, not in the similarity in content per se, but in seeing red's *having* that same content on the various occasions when subjects are in it. On this alternative, the “red” of seeing red will be an instantiated property. Although a representational state that says that things are thus-and-so need not be accompanied by any instance of things' being thus-and-so, it is itself an instance of *representing* that things are thus-and-so. But if strong representationalism is correct, this similarity is not a similarity in experience qua kind of experience it is. That is the message of the exhaustion doctrine distinctive of strong representationalism.¹³ The nature of experience qua experience is exhausted by how things are being represented to be, not by the fact that they are being so represented. But the similarity intuition that drives the knowledge argument is a view about a similarity in the nature of experience qua experience. The new similarity contention is that Mary comes to have a new kind of experience that instantiates a new property.

In sum, if strong representationalism is correct, advocates of the knowledge argument face a dilemma. If the similarity between red experiences that they see physicalists as failing to include in their picture of reality lies in the content, it implies nothing about which properties are instantiated in our world; if the similarity lies in the states with the content, it is inconsistent with the knowledge argument's claim that something about the *kind of experience* Mary has on leaving the room shows that physicalism is false.

I should highlight the fact that my argument from representationalism to the failure of the knowledge argument rests on strong representationalism. My response to Torin Alter's argument in his chapter in this volume to the conclusion that representationalism is no threat to the knowledge argument is that he successfully shows that weak representationalism is no threat to the knowledge argument. There can be different ways of representing the very same state of affairs—witness French and English sentences both representing that there is a cat before me, and the fact that formulae in polar and Cartesian coordinates can both represent a circle. Alter is right that we should distinguish the manner in which something is represented from what is represented. However, the key question for whether representationalism undermines the knowledge argument is not whether there is a content-manner distinction (there certainly is), but whether the new kind of experience Mary has when she first sees red is a reason for her to enlarge the range of properties she holds to be instantiated in our world. If, as strong

representationalism holds, the nature of the new kind of experience is exhausted by its representing as it does, it cannot provide a reason for enlarging the properties she acknowledges by the argument above: properties of how things are being represented to be are not instantiated properties; talk of properties of intentional objects is a mere manner of speech. On the other hand, if, as weak representationalism holds, the exhaustion doctrine is false, and, say, the manner in which she represents is an additional factor in making her experience the kind of experience it is, then the manner in which she represents as she does might well be a candidate to be the new similarity, the new way of categorizing items, that advocates of the knowledge argument say she learns about on leaving the room and that is left out of the physicalist scheme. And Alter will be right that espousing representationalism about experience does not buy an answer to the knowledge argument. But what he will be right about is the failure of weak representationalism to blunt the knowledge argument.

References

- Armstrong, D. (1961). *Perception and the Physical World*. London: Routledge and Kegan Paul.
- Armstrong, D. (1962). *Bodily Sensations*. London: Routledge and Kegan Paul.
- Block, N. (2003). Mental Paint. In *Reflections and Replies: The Philosophy of Tyler Burge*, ed. M. Hahn and B. Ramberg: 165–200. Cambridge: MIT Press.
- Foster, J. (2000). *The Nature of Perception*. Oxford: Oxford University Press.
-  [Link ▶ OSO X-Reference](#)
- Harman, G. (1990). The Intrinsic Quality of Experience. In *Philosophical Perspectives 4*: 31–52. Reprinted in *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzelde: 663–75. Cambridge: MIT Press, 1998.
- Hinton, J. M. (1973). *Experiences*. Oxford: Clarendon Press.
- end p.63
- Hume, D. (1739). *Treatise of Human Nature*.
- Jackson, F. (1977). *Perception*. Cambridge: Cambridge University Press.
- Jackson, F. (2004). Representation and Experience. In *Representation in Mind: New Approaches to Mental Representation*, ed. H. Clapin, P. Slezaek, and P. Staines: 107–24. Amsterdam: Elsevier.
- Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview. Revised version in *The Nature of Consciousness*, ed. by N. Block, O. Flanagan, and G. Güzelde: 597–616. Cambridge: MIT Press, 1997.
- Lycan, W. (1996). *Consciousness and Experience*. Cambridge: MIT Press.
- Moore, G. E. (1903). The Refutation of Idealism. *Mind* 12: 433–53.  [Link ▶ OUP Resource](#)
- Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge: MIT Press.
- end p.64

four Does Representationalism Undermine the Knowledge Argument?

Torin Alter

The knowledge argument aims to refute physicalism, the view that the world is entirely physical. The argument first establishes the existence of facts (or truths or information) about consciousness that are not a priori deducible from the complete physical truth, and then infers the falsity of physicalism from this lack of deducibility. Frank Jackson (1982, 1986) gave the argument its classic formulation. But now he rejects the argument (Jackson 1998b, 2003, chapter 3 of this volume). On his view, it relies on a false conception of sensory experience, which should be replaced with representationalism (also known as intentionalism), the view that phenomenal states are just representational states. And he argues that mental representation is physically explicable. I will argue that Jackson's representationalist response to the knowledge argument fails. Physicalists face a representationalist version of the knowledge argument that inherits the force of the original. Reformulating the challenge in representationalist terms does little, if anything, to help physicalists answer it.¹

Jackson's Arguments

The Knowledge Argument

You know the story. Mary is raised in a black-and-white room and has no color experiences. She learns everything in the completed science of color vision by watching lectures on black-and-white television. These lectures include “everything in completed physics, chemistry, and neurophysiology, and all there is to know about the causal and relational facts consequent upon all this, including of course

end p.65

functional roles” (Jackson 1986: 291). Then she leaves the room or is given a color television: she sees colors for the first time.

Jackson's version of the knowledge argument runs roughly as follows. Mary knows all the physical facts before she leaves the room. Intuitively, when she leaves she learns new facts about the phenomenal character of color experiences. For example, she learns what it's like to see red. These facts must be nonphysical; otherwise, she would have known them before leaving the room. Therefore, the complete physical truth cannot be (or metaphysically necessitate) the complete truth about the world: physicalism is false.²

The Argument from Representationalism

Jackson maintains that if physicalism is true, then all facts about consciousness are a priori deducible from the physical facts.³ He also maintains that the latter deducibility claim faces a serious challenge from the intuition that Mary learns new facts when she leaves the room. But now he accepts the deducibility claim and so rejects the intuition that she learns new facts. He argues that the intuition about Mary depends on a misconception about the nature of sensory experience. On his view, the correct conception is representationalism, on which

(J1) All facts about the phenomenal character of color experiences concern the representational character of these experiences.⁴

He also proposes a physicalist account of the representational character of color experiences. This leads him to the second main step of his argument:

(J2) Mary can deduce all the facts about the representational character of color experiences from the physical facts without leaving the black-and-white room.

Together, J1 and J2 entail that Mary can deduce all the facts about the phenomenal character of color experiences from the physical facts. If she can, then the knowledge argument is unsound. Call this *the argument from representationalism*.

The Argument for J1

Jackson bases his representationalist view largely on the idea that color experiences are diaphanous (or, as it is sometimes put, transparent).⁵ This idea is often
end p.66

expressed as a thesis about introspective attention: the thesis that it is impossible to attend to the phenomenal character of one's experiences except by attending to what one's experiences represent (Kind 2003, Tye 2000).

Jackson expresses the idea somewhat differently. He writes,

I start with the diaphanousness of experience: G. E. Moore's thesis that the qualitative character of experience is the character of the putative object of experience. The redness of sensings of red is the putative redness of what is seen; when vision is blurred, what is seen appears to be blurred; the location quality of a sound is the putative location of the sound; the experience of movement is the experience of something putatively moving; and so on. (2003: 257)⁶

The two formulations are not plainly equivalent, if only because Jackson's does not mention attention. But here this does not matter much. What I will say about diaphanousness applies to both formulations.

Jackson takes diaphanousness for granted, remarking, "The case for it is widely accepted and it is especially appealing in the case of our topic, colour experience" (2003: 258). He suggests that diaphanousness leaves us with two options. One is the sense data theory, on which "experiences are composed of an act of awareness directed to an object or sense datum which bears the qualities" (258). Advocates of the knowledge argument often say

that when Mary leaves the room and sees a red tomato, she learns about *phenomenal redness*, a property that determines the phenomenal character of seeing red. The sense data theorist adds that phenomenal redness is an instantiated property of a mental object, a sense datum, to which her experience is directed. On this view, it is plausible that there are facts that one cannot know without being directly acquainted with phenomenal redness: facts such as *experience E is phenomenally red*.

The other option, which Jackson prefers, is representationalism (or intentionalism). On this view, seeing red is a representational state. The inscription “red” need not be red in order to refer to that color. Likewise, an experience need not be red, phenomenally or otherwise, in order to represent the world as having something red in it. According to Jackson,

Intensionalism tells us that there is no such property [as instantiated phenomenal redness]. To suppose otherwise is to mistake an intensional property for an instantiated one. (Jackson 2003: 262)

On his view, since seeing red does not have any such property as phenomenal redness, there are no such facts as *experience E is phenomenally red*.

There are facts in the vicinity, but on Jackson's view all of them concern the experience's representational features. Some concern representational content. For example, seeing red represents something as being red. But there is more to mental representation than content. Two distinct states might represent the same content in different ways. Seeing red represents redness in a distinctive, phenomenal way, which may be absent in other cases of representing redness. As Jackson puts it, seeing red has a distinctive “feel.” Or as I shall say, seeing red *phenomenally represents* redness. Representational character includes both content and how (phenomenally or otherwise) the content is represented. On Jackson's view, representational character completely determines the nature of color experiences: on his view, there are no nonrepresentational features that play a role in determining their phenomenal character. From this, J1 follows: All facts about the phenomenal character of color experiences concern their representational character.

The Argument for J2

According to J2, Mary, while in the room, can deduce all the facts about the representational character of color experiences from the physical facts. In support of J2, Jackson notes that we already have well-developed theories that explain mental representation in physical terms: “accounts that talk of co-variation, causal connections of various kinds, selectional histories, and the like” (Jackson 2003: 263). But as he recognizes, this consideration goes only so far. Such theories may explain mental representation in general. That is, they may explain how it is that a mental state (e.g., a belief or a visual experience) might represent that the world is such-and-such—how a state might have a certain representational content. But again, in the case of color experiences, representational character includes not only content but also phenomenal representation (feel). Therefore, to complete his case for J2, Jackson must defend the view that all facts concerning phenomenal representation are a priori deducible from the physical facts.

In his 2003 essay, Jackson addresses this problem by identifying “five distinctive features of cases where our sensory experience represents that things are thus and so” (269):

First, such representation is rich. Visual experience represents how things are here and now in terms of colour, shape, location, extension, orientation, and motion... .

Secondly, it is *inextricably* rich. ... [Y]ou cannot prise the colour bit from the shape bit of a visual experience... .

Thirdly, the representation is immediate. Reading from a piece of paper that there is something of such and such a color, location, *etc.* typically induces a belief that represents that there is, but it does so *via* representing that there is a piece of paper with certain marks on it... .

Fourthly, there is a causal element in the content. Perception represents the world as interacting with us. ... Vision represents things as being located where we *see* them as being, as being at the location from which they are affecting us via our sense of sight. Finally, sensory experience plays a distinctive functional role. ... [I]t will determine a function that maps states of belief onto states of belief. A subject's posterior state of belief supervenes on their prior state of belief conjoined with their sensory experience. (269–70)

Jackson contends that these five features explain the distinctive way sensory experience represents—that is, they explain phenomenal representation. As he puts it,

If a representational state's content has inextricably and immediately the requisite richness, and if the state plays the right functional role [and has a causal element in the content], we get the phenomenology for free. (270)

end p.68

And Jackson implies that all the facts about the five features are physically explicable: these facts are a priori deducible from the physical facts.

Thus, Jackson's argument for J2 comprises three main claims:

1. *Mental representation in general is physically explicable*: all facts about mental representation in general (a feature of belief no less than sensory experience) are a priori deducible from the physical facts, perhaps via one of the leading theories currently on offer.⁷
2. *The five-feature analysis explains phenomenal representation*: the way color experiences phenomenally represent, the feel, “is a matter of immediacy, inextricability, and richness of representational content, and the right kind of functional role” (Jackson 2003: 271), plus a causal element in the content.
3. *All five features are physically explicable*: all facts about the five features (immediacy, inextricability, etc.) are a priori deducible from the physical facts.

Against the Argument from Representationalism

In response to the argument from representationalism, advocates of the knowledge argument might question representationalism or Jackson's argument for it.⁸ I will do neither. In my view, advocates of the knowledge argument can accept both. I will question Jackson's argument for J2, the premise that Mary can deduce all the facts about the representational character of color experiences from the physical facts, without leaving the room.

The problem concerns Jackson's five-feature analysis. Let us grant the other two components of his argument for J2, that both mental representation in general and all five features in particular are a priori deducible from the physical facts. Suppose that, while in the room, Mary does the deduction. So, for example, if the correct theory of how belief (or belief content) represents is a causal theory, then she knows all the relevant facts about causation and how the theory applies to these facts. Further, she knows all about the inextricable richness, the functional role, and so on, pertaining to seeing red. Now, when she leaves the room and sees a red tomato, does she learn anything new about phenomenally representing red? There is a strong intuition that she does.⁹ As a substitute for seeing red, her knowledge of the five features and the correct theory of mental representation seems hardly better than, say, her knowledge of neurobiology. Intuitively, it seems that none of this knowledge puts her in a position to deduce the distinctive, phenomenal way in which seeing red represents. How strong is the intuition? It is exactly as strong as the corresponding intuition about the Mary case before representationalism was brought onto the scene—the intuition that when she leaves the black-and-white room, she learns something about phenomenal character.

end p.69

Granted, we cannot describe Mary's epistemic progress in terms of her becoming acquainted with certain phenomenal properties, if no such properties are instantiated in her experiences. But other descriptions are available. Following David Chalmers (2004b), we could say that she learns more about the phenomenal *manner of representation* in which the color-sighted ordinarily represent redness. One could also say that seeing the tomato allows Mary to eliminate epistemic possibilities concerning how seeing red represents: possibilities that she cannot eliminate, or fully understand, before she leaves the room, despite her comprehensive physical knowledge. For example, when she sees the tomato she can eliminate the possibility that what it's like to represent redness is *G*, where *G* is what it's like to represent greenness.¹⁰ Her knowledge of the five features is third-person knowledge, which is deducible from her vast store of physical knowledge. The intuition that she acquires further first-person knowledge when she leaves the room persists. If this is right, then Jackson's five-feature analysis of phenomenal representation is inadequate.

Jackson does present one argument for his analysis. He writes:

Think of what happens when you summon up a mental image of an event described in a passage of prose. To make it image-like, you *have* to fill in the gaps; you have to include a red shirt kicking the winning goal from some part of the football field with some given trajectory, you have to make the goal scorer some putative size or other, you have to locate the goal somewhere, and so on and so forth. ... Also, you need to create a representation that represents inextricably. The “part” that delivers the size of the scorer

is also the “part” that delivers the putative location of the scorer and the colour of the shirt. And so on. To the extent that you succeed, you create a state with a phenomenology. (Jackson 2003: 270–71)

This is suspicious. Suppose Jackson is right that in order to form a certain mental image from a passage of prose, one must create a representational state that is inextricably rich, has a causal element in its content, and so on. It does not follow that knowing all about this inextricable richness and so on would allow one to a priori deduce the phenomenal manner in which the state represents. Yet the latter deducibility claim is what he needs to establish J2, the premise that Mary can deduce all the facts about the representational character of color experiences from the physical facts, without leaving the room.¹¹
end p.70

After presenting his analysis, Jackson writes, “Obviously, there is much more to say here, both by way of elucidation and by way of defense” (2003: 270). Might more elucidation and defense resolve the difficulty I have raised? I do not see how. Jackson seems to imply that his analysis need only provide a way of distinguishing between the experiences of the color-sighted and those of “the blind sighted, the believers in what is written on notes, and the bold guessers” (270). Perhaps his analysis accomplishes this. For example, blindsight experiences might differ from color experiences with respect to the richness of the representational features. But this sets the bar too low. To complete his case for J2, his analysis must do more: it must capture how color experiences phenomenally represent.¹²

This would be no mean feat. There is something it's like to represent colors in the way that Mary does not do until she leaves the room, and the idea that Jackson's analysis provides a way of deducing these phenomenal manners of representation from the physical facts gives rise to familiar intuitive doubts. We are left with the intuition that Mary acquires information about these manners of representation when she leaves the room. If she does, then the analysis leaves out something crucial about phenomenal representation. The difficulty extends beyond the specific analysis that Jackson articulates. Explaining phenomenal representation in physical terms presents the same intuitive difficulties as explaining phenomenal character in physical terms. Indeed, the issues of whether representationalism is true and whether the knowledge argument is sound would appear to be substantially, if not entirely, independent.¹³

Sometimes representationalism is understood to entail physicalism by definition. But nonphysicalists can accept representationalism in the sense of J1: they can accept that all facts about the phenomenal character of color experiences concern their representational character. Nonphysicalist representationalists will argue that certain aspects of representational character, such as phenomenal manners of representation, are not physically explicable. Further, nonphysicalists can accept the diaphanousness of experience. They can argue that, although one cannot attend to the phenomenal character of one's experiences except by attending to what one's experiences represent, experiences involve nonphysical representational properties—properties distinct from those that one's experiences represent.

Let me put my main point another way. The knowledge argument applies to all versions of physicalism. This includes the conjunction of J1 and J2, since this conjunction

constitutes a representationalist version of physicalism. Therefore, to use these claims to answer the knowledge argument would be question-begging unless independent reasons for believing them were provided—reasons that do not assume physicalism. Perhaps such reasons could be given. But then it would be these reasons, not the conjunction of J1 and J2, that answer the knowledge argument.

Jackson's five-feature analysis of phenomenal representation would, if successful, provide such a reason (assuming that all five features are physically explicable). But the doubts that the Mary case raises for familiar versions of physicalism apply
end p.71

with equal force to his analysis. So the analysis leaves physicalists back at square one: they must find a way to answer the challenge the Mary case presents. At the end of “Mind and Illusion” (2003: 271), Jackson endorses the Lewis-Nemirow ability hypothesis, on which Mary acquires abilities but no information when she leaves the room.¹⁴ This too would constitute an independent basis for rejecting the knowledge argument. But then it is the ability hypothesis, not representationalism, that answers the knowledge argument. Moral: Representationalism does not provide any clear resources for answering the knowledge argument.

Weak, Strong, and Ultrastrong Representationalism

So far, I have responded to Jackson's arguments as presented in his 2003 essay. In chapter 3 of this volume, he replies to this response by contending that my argument succeeds only if I assume an overly weak form of representationalism.¹⁵ He characterizes the strong form of representationalism that he endorses as follows:

Strong representationalism ... is the doctrine that the content of an experience *plus* the fact that the experience represents the content as obtaining in the way distinctive of perceptual representation are what determines the experience's nature without remainder. (58)

Weak representationalism is the same thesis without the “without remainder” clause: on this view, although “experiences must by their very nature represent,” their representational character is not “the sole determinant of experiential nature.”

However, my argument does not concern weak representationalism. Weak representationalism is inconsistent with J1, and I grant J1 for the sake of argument. Again, J1 says that all facts about the phenomenal character of color experiences concern their representational character. I see no difference between J1 and strong representationalism that has any relevance to my argument. In effect, strong and weak representationalism differ over whether there are nonrepresentational phenomenal properties that partly determine the nature of color experiences. The weak form allows that there may be such properties, and the strong form excludes them. But J1 excludes them, too. If there were such properties, then there would also be facts about phenomenal character that do not concern representational character. And J1 says that there are no

such facts. My main contention is that even if *strong* representationalism is true, the knowledge argument retains its force.

I do make assumptions that certain representationalists may deny. In particular, I assume that the nature of a color experience is determined at least partly by *representational* properties, such as the property of representing phenomenal redness in a certain manner, and not just by *represented* properties, such as redness. Might Jackson accept an even stronger version of representationalism with which this assumption conflicts? That would explain why he contends that I assume an overly weak form of the view. We could define *ultrastrong representationalism* as the thesis that all facts about the phenomenal character of color experiences concern the properties those experiences represent.¹⁶

But this suggestion only raises further problems. For one thing, in defending strong representationalism, Jackson seems to commit himself to the assumption that representational properties play a role in determining the nature of a color experience. Represented properties may determine the nature of an experience by figuring into its intentional content. But on strong representationalism, an experience's nature is determined by not only its content but also “the fact that the experience represents the content as obtaining in the way distinctive of perceptual representation.” Such a *way* would appear to be a representational property, not a represented property. In fact, this *way* seems to be precisely what I earlier called a phenomenal manner of representation. Moreover, the assumption that representational properties at least partly determine the nature of color experiences is plausible; and the intuition that Mary learns something upon leaving the room supports it. She can infer from what the lectures teach her that redness is instantiated in various objects outside the room. If she learns anything when she leaves and sees a red tomato, she learns how we visually represent this color. How we do this—the phenomenal manner of representation—is a representational property, not a represented one: a feature of our experiences, not tomatoes. Other creatures may represent the same color in a different phenomenal manner. Seeing the tomato helps her understand how we typically do this. This remains plausible even if we assume strong representationalism and the diaphanousness thesis (that is, even if we assume that Mary's experience has no nonrepresentational phenomenal qualities and that she cannot attend to the relevant phenomenal manner of representation without attending to the tomato's redness). And though the tomato's redness may or may not play a role in determining the nature of color experiences, the relevant phenomenal manner clearly plays a determining role.¹⁷

end p.73

But suppose one embraces an ultrastrong version of representationalism that eschews phenomenal manners of representation. On this view, color experiences involve only what Chalmers (2004b) calls “pure representational properties”: properties of representing *tout court* that such-and-such is the case. How would the ultrastrong representationalist distinguish between phenomenal and nonphenomenal representation? He would simply maintain that certain contents can be represented in experience but

cannot be represented in the absence of a relevant experience. Does this view blunt the force of the knowledge argument?

No. Granted, on ultrastrong representationalism we cannot describe what Mary learns upon leaving the room in terms of phenomenal manners. But the intuition that she learns something remains unaffected. Assuming ultrastrong representationalism, we would express the intuition as the claim that Mary comes to know about the instantiation of certain pure representational properties. The knowledge argument would then have the upshot that some pure representational properties are nonphysical—a result no physicalist can accept. So even ultrastrong representationalism fails to undermine the knowledge argument.¹⁸

Conclusion

Jackson attempts to answer the knowledge argument by combining representationalism about color experiences with a physicalist account of phenomenal representation. I have argued that this strategy cannot work. The problems that the Mary case creates for physicalist accounts of phenomenal character carry over undiminished to physicalist accounts of phenomenal representation, including the account Jackson proposes. In the debate over the knowledge argument, representationalism would appear to be a red herring.

This result suggests a more general moral: bringing representationalism to bear on the debate over whether consciousness is physical leaves everything more or less as it was. Suppose, on the one hand, that representationalism is true. Then instead of asking, “Is phenomenally conscious experience physical?” we should ask, “Is phenomenal representation physical?” But the latter question raises the same issues as the former. Physicalist accounts of phenomenal representation face representationalist versions of the antiphysicalist arguments—the knowledge argument, the conceivability argument, and so on (Chalmers 1996, 2003)—and representationalism does not seem to provide any resources for answering these arguments. Similarly, antiphysicalist representationalists face representationalist versions of the familiar problems with antiphysicalism. For example, these philosophers must contend with

end p.74

the positive arguments for physicalism, and they must explain how nonphysical aspects of mental representation relate (causally or otherwise) to physical phenomena. Suppose, on the other hand, that representationalism is false. Does this help resolve the issue of whether physicalism is true? It seems not. If we reject representationalism, then it will be natural to focus the debate on nonrepresentational aspects of conscious experience rather than on phenomenal representation. But this changes nothing of substance. The nonrepresentational aspects of experience may be nonphysical, but then again they may be physical. Perhaps it is now clear why the argument from representationalism must fail: the issues of whether representationalism and physicalism are true are orthogonal.

Acknowledgments

This chapter emerged from my commentary on Frank Jackson's presentation at the 2002 NEH Summer Institute on Consciousness and Intentionality, at the University of California, Santa Cruz. I also presented a draft of this chapter at the 2003 Pacific Division Meeting of the American Philosophical Association. For helpful comments and discussions, I thank William FitzPatrick, Frank Jackson, Amy Kind, Stuart Rachels, Daniel Stoljar, my APA commentator Robert van Gulick, Sven Walter, those who attended the presentations, and especially David Chalmers. I began work on this chapter during a leave that was made possible by an American Philosophical Society Sabbatical Fellowship and the University of Alabama. I thank both institutions.

References

- Alter, T. (1998). A Limited Defense of the Knowledge Argument. *Philosophical Studies* 90: 35–56. [Link](#)
- Block, N. (1990). Inverted Earth. In *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 53–70. Atascadero, CA: Ridgeview.
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Chalmers, D. J. (2003). Consciousness and Its Place in Nature. In *The Blackwell Guide to the Philosophy of Mind*, ed. P. Stich and T. Warfield. Oxford: Blackwell. Reprinted in *The Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 247–72. New York: Oxford University Press, 2002.
- Chalmers, D. J. (2004a). Phenomenal Concepts and the Knowledge Argument. In *There's Something about Mary*, ed. P. Ludlow, D. Stoljar, and Y. Nagasawa: 269–98. Cambridge: MIT Press.
- Chalmers, D. J. (2004b). The Representational Character of Experience. In *The Future for Philosophy*, ed. B. Leiter: 153–81. Oxford: Clarendon.
- Conee, E. (1994). Phenomenal Knowledge. *Australasian Journal of Philosophy* 72: 136–50. [Link](#)
- Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown.
- Harman, G. (1990). The Intrinsic Quality of Experience. In *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 31–52. Atascadero, CA: Ridgeview.
- Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36. [Link](#)
- Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy* 83: 291–95. [Link](#)
- Jackson, F. (1994). Finding the Mind in the Natural World. In *Philosophy and the Cognitive Sciences*, ed. R. Casati, B. Smith, and G. White. London: Routledge. Reprinted end p.75

in *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 162–70. New York: Oxford University Press, 2002.

Jackson, F. (1995). Postscript. In *Contemporary Materialism*, ed. P. Moser and J. Trout: 184–89. London: Routledge.

Jackson, F. (1998a). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Clarendon.

Jackson, F. (1998b). Postscript on Qualia. In *Mind, Method, and Conditionals: Selected Essays*: 76–79. London: Routledge.

Jackson, F. (2003). Mind and Illusion. In *Minds and Persons: Royal Institute of Philosophy Supplement* 53, ed. A. O'Hear: 251–71. Cambridge: Cambridge University Press.

Jackson, F. (2004). Forward: Looking Back on the Knowledge Argument. In *There's Something about Mary*, ed. P. Ludlow, D. Stoljar, and Y. Nagasawa: xv–ix. Cambridge, MA: MIT Press.

Kind, A. (2003). What's So Transparent about Transparency? *Philosophical Studies* 115: 225–44. [!\[\]\(a870788d6ed9b8fd294b7654a8c8526b_img.jpg\) Link](#)

Lewis, D. (1988). What Experience Teaches. *Proceedings of Russellian Society (University of Sydney)*. Reprinted in *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 281–94. New York: Oxford University Press, 2002.

Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives* 4: *Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview. Revised version in *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzelde: 597–616. Cambridge: MIT Press, 1997.

Moore, G. E. (1903). The Refutation of Idealism. *Mind* 12. Reprinted in G. E. Moore, *Philosophical Studies*, 1922/1965: 1–30. London: Routledge and Kegan Paul.

Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. In *Mind and Cognition*, ed. W. Lycan: 490–99. Cambridge: Basil Blackwell.

Robinson, W. (2003). Jackson's Apostasy. In *Philosophical Studies* 111: 277–93.

[!\[\]\(9c2e8d1b5bd77cb5c9f83b7a9cff79fd_img.jpg\) Link](#)

Searle, J. (1983). *Intentionality*. New York: Cambridge University Press.

Siewert, C. (1998). *The Significance of Consciousness*. Princeton, N.J.: Princeton University Press.

Stoljar, D. (forthcoming). Consequence of Intentionalism. *Erkenntnis*.

Stoljar, D., and Nagasawa, Y. (2004). Introduction to *There's Something about Mary*, ed. P. Ludlow, D. Stoljar, and Y. Nagasawa: 1–36. Cambridge: MIT Press.

Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge: MIT Press.

end p.76

five What Is This Thing You Call Color Can a Totally Color-Blind Person Know about Color?

Knut Nordby

Let us make the following thought experiment: A boy, let us call him Gus, is raised from infancy to the age of 21 on very bland food and not allowed to eat or taste anything spicy

or strong tasting (while maintaining a nutritionally balanced diet containing all important proteins, vitamins, and minerals). However, he is allowed and encouraged to read everything about food and to discuss anything related to food and gastronomy. On his twenty-first birthday, he is taken to ethnic restaurants and treated to various dishes, such as hot curry, hot chili peppers, or wasabi (Japanese horseradish). Will he, after the shock, be able to taste the spicy flavors, and, more important, will he be able to identify the various tastes solely on the basis of his previously acquired knowledge, and without the aid of vision?

This “Porridge Paradigm” is a kind-hearted paraphrase on Frank Jackson's (1982, 1986) “Mary Contention.” Mary is similarly raised, but in a monochrome (black/gray/white) environment where she is denied all visual experience of color hues, but is encouraged to read and learn all there is about color. When, on her twenty-first birthday, she is let out into the world of full colors, will she be able to experience the color hues and identify them on the basis of her acquired knowledge?

I believe that Mary will be able to sense and discriminate color hues but will not be able to name them on the basis of her knowledge. To develop my argument, it will be necessary to look into some basic concepts of color vision as well as to look into some analogous experiences in other sense modalities.

Visual information is mediated by the sensory cells (cones and rods) of the retina of the eye. The *cones* mediate vision under bright (*photopic*) light conditions, and the *rods* mediate visual information under low (*scotopic*) light conditions. There is an intermediary (*mesopic*) region where the cones start to function. There are three types of cones, each type with its peak sensitivity to light of different wavelengths. For the sake of simplicity, we can call them *blue* (short wavelength) sensitive; *green* (middle wavelength) sensitive, and *red* (long wavelength) sensitive cones. The relative contributions of the three cone types, when combined in the color centers of the visual cortex of the brain, give rise to the sensation of color. There is one kind of rod; rods are most sensitive for middle wavelengths, but they do not contribute specifically to color perception (for details, see Rodieck 1998).

The great nineteenth-century Scottish physicist James Clark Maxwell claimed that color was an integral property of object surfaces. Thus, if you could not sense color, you would not be able to perceive the form of objects. Maxwell's maxim may not hold for chromatic colors because there are people who lack all color sensation (a condition called *achromatopsia*, or *congenital typical rod monochromacy*), but who can perfectly well see the form of objects.

I happen to be one of these people (Nordby 1990). I may thus be regarded as a living embodiment of the Mary in the gray-room thought experiment. But there are some important differences. (1) I was born without retinal cones, (2) I have always been exposed to colors, and (3) I did not, and will not, experience the “coming out” on my twenty-first birthday.

People who are totally total color-blind or achromatopic have visual input only from retinal rods and see only in contrasts of brightness. They thus live in a perpetual state of “night vision” (Sharpe and Nordby 1990a, 1990b). Typical achromatopsia (or rod monochromacy) is congenital and nonprogressive. It is characterized by:

- Total lack of color discrimination (though people with this condition can match any color hue to any other color hue or shade of gray on the basis of brightness)
- Hypersensitivity to bright light (called “photophobia”)
- Low visual acuity (typically 6/60 [or 20/200] Snellen, which means that people with this condition can resolve fine detail at 6 meters [20 feet] that people with normal acuity can resolve at 60 meters [200 feet])
- Nystagmus (involuntary, rapid side-to-side movement of the eyes; this diminishes with age)

There is an old proverb saying that: *At night, all cats are gray*, implying that people with normal color vision (*trichromates*) will experience loss of color perception at very low (scotopic) levels of illumination: night vision is achromatic.

Although typical achromatopsia is very rare (estimated at 1 in 30,000 or 1 in 40,000) there is a much rarer *cortical* form of achromatopsia. Oliver Sacks and Robert Wasserman report the story of the painter Jonathan Isacson, who lost all color perception as a result of a cerebral trauma after a car accident (Sacks and Wasserman 1987, Sacks 1995). After a couple of months, he could not even remember what colors looked like and knew only as a piece of fact that he had once been able to experience colors.

There are also *incomplete* forms of achromatopsia, as with people who are “totally color blind” but who can still see some colors.¹ These cases must not be confused with *dichromates* or *anomalous trichromates*, the traditionally “color-blind,” who are, for example, *protanopes* (red blind) or *deutanopes* (green blind).

People living under normally bright (photopic) lighting conditions will have their visual information mediated by the retinal cones and rods regardless of what wavelengths they are subjected to. Mary's black/gray/white world will thus
end p.78

stimulate her normal color-sensing system—just as black, gray, and white can be perfectly rendered on color film or by color television—so, eventually, when she is subjected to all the colors out in the real world, her visual system should be fully able to mediate light wavelength information that gives rise to color sensations. Still, of course, Mary will not know what to call the various color sensations unless she makes use of noncolor information; for example, knowing that a rose is red, she may recognize the form of a rose and deduce that therefore its color must be red. Achromatopic people do this all the time by memorizing such facts as that fire engines are red, violets are blue, grass is green, lemons are yellow, and so on. This enables them to “know” colors without experiencing the color hues. This would therefore pose a problem in testing Mary's color perception. A totally color-blind person will happily match any color hue to any other color hue or gray tone by adjusting the brightness, but Mary would not be able to do this; she will not be able to match green with a red color or gray tone, no matter what brightness they have.

A totally different situation would arise if Mary were raised under very low (scotopic) levels of illumination. Then her photopic (cone-mediated) visual system would not receive adequate input and would most likely deteriorate under lack of appropriate

stimulation. She would then probably show the signs of what was once called “miners' nystagmus,” a visual condition regularly reported in people (not least children) who for many years, often from an early age, worked in total, or nearly total, darkness in the mines. Up to the end of the nineteenth century, it was common to send children as young as six into the mines to work long hours performing some simple task, such as opening a gate for mine trucks or pulling a lever to tip the trucks. Their tasks were not considered important enough to warrant the use of lighting, which was expensive, and they usually had to work up to 14 hours a day in total or nearly total darkness. In addition, it was normally dark outside when they entered the mine, and it was often dark when they came up. The most telling symptoms in these people were nystagmus and photophobia (a misleading term because the condition has nothing to do with the irrational phobias of psychiatry). It is not known exactly how their color vision was affected, but there is some evidence that it was reduced; it is astonishing how little attention was paid to color-vision defects before the twentieth century. Thus, if Mary were raised under scotopic conditions, she would almost certainly display the typical symptoms of miners' nystagmus and would probably not be able to perceive colors, but she might be able to name the colors of known objects from general knowledge.

So What Is This Thing You Call Color?

Can an achromatopic person ever have any idea what a color experience is? Most achromatopic people think of color as some curious property of surfaces that for them is somehow related to their apparent brightness. Thus yellow looks lighter than other colors, and red looks darker.

To be able to cope in the world of color-sighted people and avoid embarrassment, most achromatopic people teach themselves the colors of common objects
end p.79

and the cultural “meaning” of some colors (e.g., red for danger, green for clear, blue for sadness).

In 1994, in the company of neurologist and writer Oliver Sacks and ophthalmologist Robert Wasserman, I visited the island of Pingelap, a small isolated atoll in the Carolines archipelago in the Pacific on which there is a high incidence of achromatopic people (Sacks 1996). We made a rather interesting observation: while testing the color-naming abilities of the islanders, we were made aware by our interpreter that there was no proper name for the color *orange* in the Pingelapese language, probably because this color was not very prominent in the local flora or fauna. However, we could easily have been fooled because those islanders who had color vision called it the “orange color,” meaning the color of oranges, which they had only recently come to know (no oranges grow on the atoll). Thus, even though the inhabitants of Pingelap traditionally had not been subjected much to the color orange, and therefore had no local name for it, they could still clearly recognize it and distinguish it from other colors by likening it to the color of a fruit they had recently been introduced to. This is as close as I have ever come to Jackson's “Mary

Contention” in real life. There is of course the possibility that the inhabitants of Pingelap had always known the color orange but that it was so irrelevant for them that they never bothered to give it a name.

One way for me to attempt to visualize the special quality of experiencing color is to liken color to the musical quality of tones, or *chroma*. Whereas colors have brightness and hue (where brightness is a function of the number of photons striking the retina, and hue is related to the wavelength of the light), tones have loudness, pitch, timbre, and chroma. A tone's loudness is a function of the sound pressure in micro pascal (μPa); a tone's pitch is a direct function of its frequency in hertz (Hz); a tone's timbre derives from the number and strength of its harmonics; and a tone's chrome (or musicality) is a function of its frequency in Hz and a cyclical element repeating each octave (a musical octave is a doubling of its fundamental frequency). When speaking of “tones,” I refer specifically to the diatonic tone scales used in Western music.

Chroma is the special property of tones that give them their *musical tonality*. Tones that are one or more octaves apart sound more similar than tones that lie close together on the scale; thus the tones middle C, c, c^1 , and c^2 (each separated from the next by an octave) have a musical property in common that is not present between C and C-sharp or between C and C-flat, the closest neighbors on the frequency scale (see fig. 5.1).

The same holds for all the other tones on the diatonic scale; thus C#, D, D#, E, F, F#, G, G#, A, A#, and B each sound more similar to tones one or more octaves above (or below) than to other tones on the scale. However, when we go above 4,000 Hz (close to the uppermost key on the piano keyboard), the sensation of musical chroma disappears, and we become *tone deaf*. This may be likened to the disappearance of color under scotopic conditions. There are people who cannot make out a melody but have perfect pitch discrimination, and we call them “tone deaf.”

In an unpublished experiment I performed some thirty years ago, I let subjects listen to short melodies played both in the 2,000–4,000 Hz octave and in the

end p.80

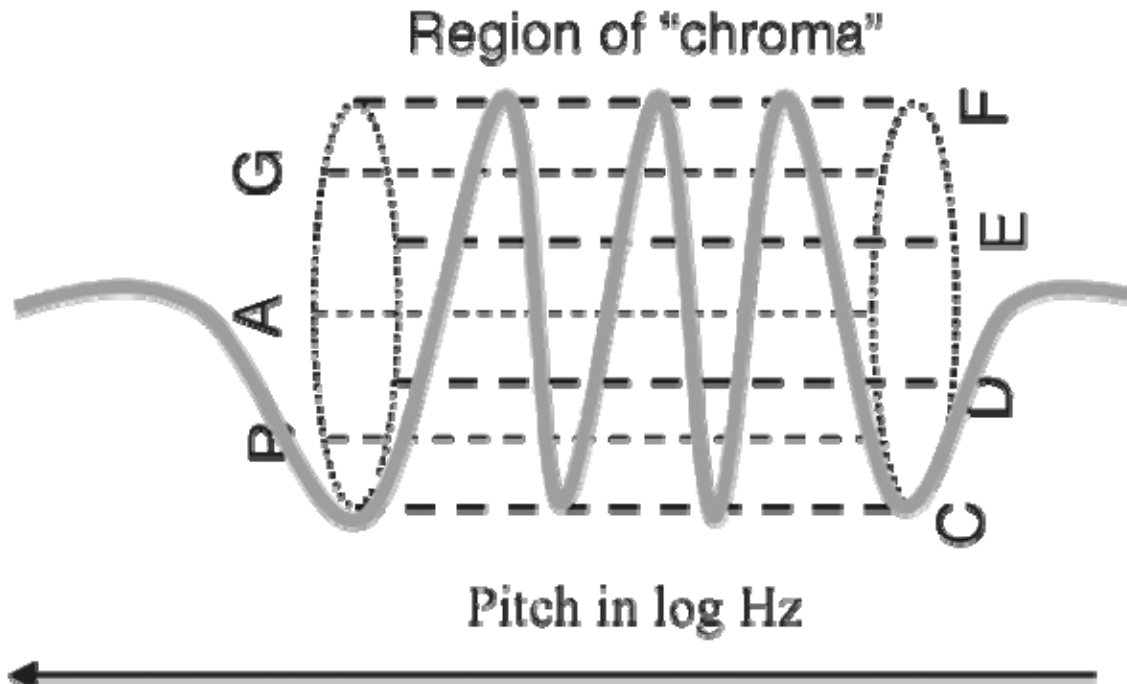


Figure 5.1. The figure depicts the cyclical tone property of “chroma” (spiral) as a function of log frequency in Hz. The tones of the diatonic scale are indicated as lines on the periphery of the cylinder. Each turn of the “chroma” line is one octave.

4,000–8,000 Hz octave. The melodies had the same rhythm and similar up-and-down movements to avoid recognition on these criteria. All the subjects could easily recognize all the melodies in the lower octave, but none could positively identify any melody in the upper octave; they had to resort to guessing. This implies that it is not possible to make music with tones that have their basic frequencies in the region above 4 kHz, and people with normal hearing thus become “tone deaf” above 4 kHz.

Some will probably object to my comparison of color to tone chroma, and they instead will propose that color is more comparable to the *timbre* or *tone color* of instruments than to the cyclical tone chroma. Timbre results from the combination of the fundamental frequency of the tone and the harmonics (or overtones) of different frequencies that produce the characteristic tone colors of instruments and makes it possible to distinguish a flute from a clarinet, a piano from a guitar, and one human voice from another. For me, timbre is a tone quality that is not dependent on its fundamental pitch, and people can be tone deaf for all timbres: timbre is a direct property of the sound, whereas chroma is a more elusive property. You may describe sound timbre by likening it to other sounds (e.g., a flute sounds like a certain bird, and a kettledrum sounds like distant thunder), but you cannot compare the quality of C-ness with D-ness, E-ness or G-ness, just as you cannot equate the color quality of *blueness* with *greenness* or *redness*.

For me, tone chroma is an inherent property of “tones” that cannot be separated from them, just as Maxwell maintained that color was an intrinsic property of object surfaces. Although the “chroma metaphor” may convey the idea of a special sensory
end p.81

property solely as an abstract thought exercise, it can never depict the actual experience of color. Colors, like tones and tastes, are firsthand sensory experiences, and no amount of acquired theoretical knowledge can create this experience. Or is this so?

The color-blind painter, Jonathan Isacson, very soon forgot about his earlier color experiences following his traumatic loss of color perception, implying that the color centers in his brain lost their original function. Most people who lose a sense modality soon forget about the sensations of the lost modality, especially if the loss occurs early in life. However, there are cases of people who have lost their vision or hearing but who have retained or even developed a very strong inner visual or auditory imagery. In hearing, the case of Beethoven comes to mind; he composed some of his finest music after going completely deaf. And there are cases of people who have lost their vision but actively developed very intense inner visual imagery; Sacks (2003) has described several such cases.

Could it be that there are some unknown people out there who cannot sense colors but who have developed an inner “color vision” of their own making? Whatever inner sensations they call “color” may have no relation to or bear no resemblance to what the color-sighted call color. However, such inner color would not help these people avoid unripe fruit or recognize objects on the basis of their hue, though it could be psychologically satisfying for their inner life.

Now that Jackson has retracted his famous *knowledge argument* (see Jackson, chap. 3, this volume; and Alter, chap. 4, this volume), it may look like flogging a dead horse to further argue against the speculative and unsubstantiated “Mary Contention,” but poor Mary has already caught the public imagination, and like many other unsubstantiated theories before it, it will take years for it to disappear. Since, as a color-blind person, I am often confronted with Jackson's contention, I appreciate this opportunity to state my view on the issue. I also wish to offer Jackson my apologies for paraphrasing his “Mary Contention,” but the temptation was just irresistible.²

References

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36.



Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy* 83: 291–95.



Nordby, K. (1990). Vision in a Complete Achromat: A Personal Account. In *Night Vision: Basic, Clinical and Applied Aspects*, ed. R. F. Hess, L. T. Sharpe, and K. Nordby: 290–315. Cambridge: Cambridge University Press.

Rodieck, R. W. (1998). *The First Steps in Seeing*. Sunderland, MA: Sinauer.

Sacks, O. (1995). The Case of the Colour-Blind Painter. In *An Anthropologist on Mars*: 1–38. New York: Vintage Books.

Sacks, O. (1996). *The Island of the Colorblind*. New York: Vintage Books.

Sacks, O. (2003). The Mind's Eye: What the Blind See. *New Yorker*, July 28.

Sacks, O., and Wasserman, R. (1987). The Case of the Color-Blind Painter. *New York Review of Books*, November 19, 1987.

Sharpe, L. T., and Nordby, K. (1990a). The Photoreceptors of an Achromat. In *Night Vision: Basic, Clinical, and Applied Aspects*, ed. R. F. Hess, L. T. Sharpe, and K. Nordby: 335–89. Cambridge: Cambridge University Press.

Sharpe, L. T., and Nordby, K. (1990b). Total Colorblindness: An Introduction. In *Night Vision: Basic, Clinical and Applied Aspects*, ed. R. F. Hess, L. T. Sharpe, and K. Nordby: 253–89. Cambridge: Cambridge University Press.

end p.83

end p.84

Part two Phenomenal Concepts

end p.85

end p.86

six What Is a Phenomenal Concept?

Janet Levin

An increasing number of physicalists agree that phenomenal concepts should be treated as special sorts of representations, conceptually independent of physical or functional descriptions, which a subject can acquire only by having the experiences they denote.¹ This account is compatible with physicalism, they argue, so long as these concepts pick out their referents *directly*, much like demonstratives, without mediation by any mode of presentation. Indeed, these physicalists stress, such a view can provide a one-stop solution to a number of well-known problems that have threatened the identification of phenomenal and physical properties: it can explain, without appeal to nonphysical properties, what Mary learns when she leaves her black-and-white room (Jackson 1982), why zombies are conceivable (Chalmers 1996), how irreducibly phenomenal concepts can refer to physical properties (White 1986), and why phenomenal-physical identities seem to open an “explanatory gap” (Levine 1983) not encountered in other cases of inter-theoretic reduction.

But there has also been increasing criticism of this account of phenomenal concepts: not surprisingly by dualists, but also by physicalists, including some who had previously endorsed the view.² My aim here is to present and evaluate these criticisms and to provide a limited defense of the demonstrative account. I do this not as an originator or early proponent of the view (or even as a current true believer), but rather as a physicalist who has argued (1991, 2002) that phenomenal concepts must be at least partially functional to provide a satisfactory response to the antiphysicalist concerns. For reasons I'll make clear in this chapter, I suspect that my defense of (what I'll call) quasi-

functionalism may have sprung from (what Freud called) the narcissism of small differences, and I now think that a demonstrative view may be able to meet my own previous worries and those alluded to above. Equally important, I'm alarmed by the emendations to the demonstrative account that have recently been suggested by some physicalists (Balog 2002; Block, chap. 12, this volume; Papineau 2002) and agnostics (Levine, chap. 8, this volume). These suggestions seem to concede too much to the antiphysicalists while accomplishing too little, and one aim of this chapter is to urge demonstrative theorists to return to their roots. Even so, my defense of the demonstrative account will be limited, because there are other, freshly articulated, worries about physicalism that suggest that it may need refinement: for example, the "harder problem of conscious experience" recently noted by Ned Block (2002).³ But a clear-eyed appraisal of its strengths is important for determining how physicalists can best proceed from here.

The Demonstrative Account

Most physicalists who take phenomenal concepts to function like demonstratives suggest that they denote whichever neural properties are causally responsible for our application of these concepts in various introspective tasks. Which tasks these are depends on whether one is interested in *type* or *token* phenomenal concepts. Token phenomenal concepts are those that can be used to pick out an *instance* of an experience with some salient qualitative character (Tye 1995), and they are taken to denote in the manner of *token-demonstrative* concepts. That is, whichever (neural) particular causes me to make introspective note of some experience I'm now having counts as the denotation of the token-demonstrative "that (experience I'm having now)." ⁴ And if I'm having more than one experience at a time (as when I simultaneously have a pain, hear a noise, and see red, or when I focus successively on two different shades of red), then I can denote distinct neural-particulars, respectively, as "this," "that," and so on, as long as I can discriminate among these experiences and successively direct my introspective attention to them at that time. Though token-demonstrative concepts can be invoked to explain certain features of introspective knowledge (Tye 1995), they don't have to be stored in memory, and their referential success implies nothing about the user's ability to recognize other such experiences as experiences of that kind. Thus their usefulness is limited, since—for reasons I'll discuss shortly—they can function only to pick out instances of experiences with phenomenal properties, and not those properties themselves.

end p.88

Of greater use against the antiphysicalist arguments are *type-demonstrative* phenomenal concepts ("that *kind* [of experience]"), which purport to pick out *kinds* or *properties* of experiences from an introspective perspective. The denotation of a phenomenal type-demonstrative will be the property—presumably physical—that's causally responsible for the application of that concept in the introspective *recognition* or *reidentification* of an experience as "that (kind) again" or "another of those."⁵ These concepts, as noted before, are taken to refer "directly"; that is, to have *no* reference-fixing "modes of presentation" or Kaplanian "characters" that can change reference from world to world (e.g., "the property I am ostending now and am disposed to identify as 'another one of those' "). Rather, their references are determined solely by the causal and dispositional relations an individual has to her internal states that are effected by an introspective "pointing in";

that is, by the *fact* that she's in causal contact with a certain property and is disposed to reidentify it on subsequent occasions.⁶ Brian Loar was perhaps the first to propose and articulate this suggestion (1990/97), and it has been adopted by many physicalists since.⁷ Loar himself characterizes these concepts as “recognitional/demonstrative,” presumably because of the crucial role that recognitional abilities play in their individuation, but I'll continue to use the term “type-demonstrative” in what follows.

I've just suggested that only concepts that fulfill the conditions for being *type-* (but not *token-*) demonstratives can denote repeatable phenomenal properties, rather than merely particular experiences that have them. But why, one might wonder, must this be so; why couldn't (what I've called) a token-demonstrative denote a phenomenal property by picking out (that is, by being caused by) an experience with that property that is attended to in introspection? The reason is that the ability to recognize or reidentify is required to underwrite *determinate reference* to a particular property. The best way—perhaps the only physicalistically acceptable way—to determine whether someone's current “pointing in” denotes what it's like to see some particular shade of red, or a more coarse-grained phenomenal property (e.g., red in general, or color), or one of a number of phenomenal properties that are instantiated in an experience but impossible to attend to selectively at that time (such as what it's like to see a square and what it's like to see red when one is looking at a red square), is to see what she is disposed to identify as *other* instances of that property. Thus I'll focus primarily on phenomenal concepts as type-demonstratives in this chapter.

end p.89

Even so, phenomenal concepts can't be fully understood on the model of nonphenomenal type-demonstrative concepts, such as “that (kind of cactus)” or “this (style of architecture),” because they are supposed to be acquired “via introspection,” or “from an introspective perspective,” and more has to be said about what exactly that entails.⁸ But there are two essential features of demonstratives that phenomenal concepts, on this view, are taken to share: first, they pick out their referents from a particular point of view—the perspective of the demonstrator—and thus are not equivalent to any nonperspectival (discursive, objective) concepts; second, they can pick out their referents “directly,” without need of identifying modes of presentation.

These essential features, many physicalists argue, serve to explain, or explain away, all the well-known phenomena that are thought to raise problems for physicalism. If phenomenal concepts are not equivalent to any physical or functional concepts, then what Mary can be said to gain when she leaves the black-and-white room is not access to a set of irreducibly phenomenal properties of color vision, but just a new way of conceptualizing the physical and functional properties that she knew about, under different descriptions, before.⁹ And just as my ability, under certain circumstances, to wonder whether *that* (pointing at myself) is me does not suggest that there's a possible world in which I am not myself, the conceivability of a zombie—that is, a molecular duplicate of me that does not feel like *that* (pointing in)—will provide no evidence that such a thing is possible. Further (in response to the “distinct property argument”), if phenomenal concepts refer “directly,” then their nonequivalence to physical or functional concepts does not imply that they must denote by means of a nonphysical mode of

presentation. And finally (in response to worries about an explanatory gap), if phenomenal concepts function like demonstratives, then phenomenal-physical identifications involving these concepts should be expected to be very different from the usual cases of inter-theoretical reduction, and thus their explanatory asymmetries need not saddle the physicalist with any unusual burden of proof.

One person's one-stop solution, however, can be another's blunt instrument, and many philosophers have charged that the strategy of treating phenomenal concepts as special sorts of type-demonstratives will ultimately fail because they are invoked for what seem to be incompatible tasks: To avoid the conclusion that phenomenal concepts denote or express irreducible phenomenal properties, the worry goes, they must denote “directly,” like other demonstrative concepts, without need for special, arguably nonphysical, modes of presentation. But if these concepts are sufficiently “thin” to denote the way a demonstrative does—that is, by serving merely as a pointer directed at (that is, differentially caused by) a type of experience—then they are insufficiently robust to account for what seems special about phenomenal concepts, or why the knowledge that Mary acquires when she leaves her black-and-white room seems so substantive, or why the explanatory gap appears just in cases of physical-phenomenal identification. Let us address these worries in turn.

end p.90

Are Phenomenal Demonstratives “Empty”?

There's a simple and basic objection to the demonstrative model that has been explicitly expressed by a number of theorists, and it may well lie behind what seems to be increasing discomfort with the view. These critics acknowledge that phenomenal demonstratives may be different from other demonstratives (e.g., “*this* [directed at a chair or table],” “*that* [style of architecture]”) in that they are acquired by means of introspective attention to one's own experiences. But, they note, that alone can't mark the difference between phenomenal concepts and others because there are other, *nonphenomenal*, demonstratives that can be acquired and deployed from an “introspective perspective” as well. For example, Horgan and Tienson (2001) and Block (in conversation) have suggested that blindsight subjects could equally well use demonstratives in introspection to denote their peculiar, nonphenomenal, states.¹⁰ Thus, the objection continues, to characterize phenomenal concepts as introspectively acquired demonstratives doesn't capture what's special about them.

To stanch these and related worries, some physicalists (and agnostics) have attempted to embellish the “classic” account of demonstrative concepts, suggesting that phenomenal concepts must involve “acquaintance” with phenomenal properties (Levine, chap. 8, this volume), or be “partially constituted” by phenomenal qualities (Block 2002), or “quote” those properties (Papineau 2002, Balog 2002) by containing instances or tokens of them, or have some other relation to the qualities they denote besides simply being differentially *caused* by them in the appropriate circumstances, or producing dispositions to reidentify them when they occur again.

But these proposals are obscure and, if spelled out in a way acceptable to physicalists, may raise analogous questions: Why couldn't a blindsight subject's introspectively deployed type-demonstratives quote, or be partially constituted by, the properties *they* denote as well? More important, these proposals are unnecessary because they give undue credence to intuitions that can be explained away by
end p.91

clear-eyed adherence to the “classic” account of demonstratives, and unflinching acceptance of what physicalism entails.

In particular, if phenomenal concepts really function like introspectively deployed demonstratives, then all that's needed to distinguish them from introspectively deployed nonphenomenal demonstratives are differences in what they denote. This is the standard way to think about demonstratives deployed from similar perspectives: if my (token) “that” (pointing out the window) picks out a car with a particular constellation of properties, then that's what determines whether I've denoted a Maserati (and not the nearby Ford), and if my “that (kind)” consistently picks out all and only Maseratis, then I've managed to denote the (very distinctive) property of being a Maserati (rather than being a Ford). My demonstratives may thus be regarded as special (after all, they're *Maserati* demonstratives), but this is entirely because of the features of what they denote. Similarly (focusing here on token-demonstratives), the states a blindsight subject's demonstratives denote are not experiences, since these states (by hypothesis) have no phenomenal properties, whereas the states denoted by a normal person's demonstratives do have phenomenal properties. If phenomenal properties are identical with physical properties, then there will be physical differences between our experiences and the introspectively denoted neural states of a blindsight subject. Some of these physical properties, of course, will be “felt” by the subjects who have them, and some will not—but that's just what it is for some, but not all, of them to be *phenomenal*. And that's all it should take for the demonstratives that denote these states, or their associated phenomenal properties, to be “special” as well.¹¹

What makes this view seem problematic, I suspect, is the thought that someone looking through a scanner at the brains of a normal subject and a blindsight subject might not realize (without consulting her textbook) that the neural states (or properties) denoted by the normal subject's demonstratives have a special phenomenal “feel.” But this is no objection to the demonstrative account, which is intended to *acknowledge* the conceivability of physical but not phenomenal duplicates, and the existence of an explanatory gap, while explaining how phenomenal properties can nonetheless be identical with physical properties. To insist that one be able to tell immediately which states, and thus which demonstratives, are phenomenal is to deny a premise shared by demonstrative theorists and dualists: namely, that one can't read off the phenomenal character of mental states from their physical or functional descriptions.¹²

Still, there are other questions about the view that must be given serious attention. In the next two sections, I'll try to dispel the worry I have had about the demonstrative theory, namely, that physicalists ought to be able to do better in explaining (what seems to be) the rich and robust knowledge that Mary gains when she leaves her black-and-white room. I'll also address a separate worry about the view (Raffman 1995) that may seem

even more basic—and damaging. In the remainder of the chapter, I'll address the question of how the account fares in the “two-dimensional semantics” as elaborated by David Chalmers and Frank Jackson; in particular, I'll explore the question of whether treating phenomenal concepts as demonstratives captures the way we take the contents of these concepts to be determined in various possible worlds. Finally, I'll consider whether the view can deal satisfactorily with a refinement of the “distinct property argument” proposed recently by Stephen White (1999, this volume, chap. 11).

Phenomenal Demonstratives and Knowing What It's Like

The first question is whether the knowledge Mary gains from experience can be adequately explained as the acquisition of phenomenal demonstratives that can figure in propositional knowledge. The worry I expressed before, on behalf of the critics, was whether the new facts expressed with these concepts are too “thin” to account for what seems to be the rich and robust knowledge of experience that Mary gains when she leaves her black-and-white room. This, of course, is a serious worry. There is another worry as well, however, that seems to be more basic, namely, whether a type-demonstrative account can even begin to do the job.

This worry is highlighted in a recent article by Diana Raffman (1995), who cites empirical studies that show that we're incapable of having enough type-demonstrative concepts to account even minimally for the knowledge we have of our own phenomenal states. Raffman takes these studies to raise insoluble problems for the demonstrative account of knowing what it's like—and therefore (on the assumption that

end p.93

this is indeed the best physicalist response to the knowledge argument) for physicalism itself.¹³ What I'll argue, though, is that there is a way of understanding *knowing what it's like* that permits Raffman's observations to be accommodated by physicalists—and, in addition, that helps to explain how the acquisition of phenomenal demonstratives can be seen to provide knowledge as rich and robust as our intuitions demand.

Raffman's worry is this: A number of psychological studies suggest that normal subjects can discriminate (that is, discern just noticeable differences [jnd's] among) far more shades of color than they can reidentify over time. For example (her example), subjects can discriminate subtly different shades of red (they distinguish, say, red-31 from red-32 and red-33) when presented with them simultaneously, but can't consistently pick out red-31 (sometimes choosing red-32 and red-33) as “that (color I just saw)” when the various shades are presented one by one. The same will be true, presumably, when subjects attempt to reidentify the qualitative properties of their color *experiences*.

Thus, if having a phenomenal concept of a certain experiential property requires the ability to reidentify the property when *encountering it by itself* (as well as being able to discriminate it from others), then we can't have phenomenal concepts of such finely individuated shades as red-31 and red-32. We may have discursive or “theoretical” concepts of them, concepts such as “the shade that normal subjects judge to be three jnd

units from unique red,” and we may have coarser-grained phenomenal concepts, which permit the reidentification of shades in a broader band.¹⁴ In addition, we can use a *token* phenomenal concept to pick out an *instance* of an experience with that phenomenal property just by attending to it in introspection. But the range of our type-demonstrative concepts of color experience falls short of the range of color experiences we can discriminate in introspection (when they're simultaneously displayed).

So, the argument continues, if coming to know what it's like to have an experience is to acquire a type-demonstrative concept of it, then there are many experiences—for example, the experience of red-31—that we *can't know what it's like to have*. But clearly a person who is currently discriminating red-31 from other shades in introspection *will* have this knowledge: it's perfectly intuitive to think that if Mary were presented with a sample of red-31 upon leaving her black-and-white room, she'd describe herself as now (finally!) knowing what it's like to see red-31. Thus, the suggestion that what Mary gains from experience are new type-demonstrative concepts derived from introspection cannot account for all that Mary comes to know.

Physicalists, however, can respond to these worries by suggesting that, contrary to a crucial premise in the argument, Mary *doesn't* know what it's like to see
end p.94

red-31, even though she can discriminate a red-31 shade from other shades in simultaneous presentation.¹⁵ Or, equivalently but more diplomatically, physicalists could claim that our notion of knowing what it's like is ambiguous.¹⁶ In one sense, anyone who can discriminate the experience of some particular shade—say, red-31—from others in introspection counts as knowing what it's like to experience that shade. All that this sort of knowledge requires is the ability to apply, successively in introspection, a *token-demonstrative* concept to a particular experience of red-31, and then to a different experience. As long as there's a physical difference between these experience-tokens, Mary can have different token-demonstrative concepts of them. So Mary does, in this sense, know what it's like to see red-31 when she's looking at, and attending to, a sample of that shade.

But in another, more substantive sense, those who are not disposed to reidentify the experience of a shade as fine-grained as red-31, or to recognize instances of it when it occurs by itself in introspection, do *not* know what it's like to see red-31 (rather than, say, what it's like to see an instance of a broader band of the red spectrum that includes red-31). These abilities are required for having (first-person) knowledge of a phenomenal type, for without them, there is no way to determine *which* property is being demonstrated by a subject who “points in” and thinks, “This is what it's like to see red.” In *this* sense, of course, Mary may not know what it's like to see red-31 even after she leaves her black-and-white room and stares, attentively, at a rose of that color. And if Raffman's facts are correct, *neither may we*.¹⁷ But why should this be a problem for physicalism, rather than merely a fact about human memory and categorization capacities?

Antiphysicalists may argue that it's “just intuitively clear” that, in this situation, Mary *does* know what it's like to see red-31 in the second, more substantive sense. And their argument may seem to be supported by another psychological fact adduced

end p.95

by Raffman, namely, that our ability to reidentify colors (and thus, presumably, color experiences) seems to exhibit certain asymmetries. One can't, Raffman notes, simply claim that though we have phenomenal concepts of a fairly broad band of phenomenal reds and phenomenal greens—or maybe even phenomenal indigos and chartreuses—we just don't have concepts as fine-grained as red-31 or red-32. For it turns out that people are in fact very good at reidentifying—not just discriminating in simultaneous presentation—the “unique” shades of red, green, yellow, and blue, even though they occupy an equally fine-grained place in our color quality space. (“Unique” colors are those that we might call “pure,” e.g., a red with no yellow or blue in it at all.) So one might imagine Mary, just out of her black-and-white room, being presented with two roses, one that's red-31 and one that's unique red. As she's looking at them and reflecting on her new experiences, it may seem that she has equal grasp of their phenomenal properties. But, on the view that takes knowledge of what it's like to see colors as the possession of type-demonstrative concepts of color experience, Mary can have robust, substantive knowledge only of what it's like to see unique red. And this may seem counterintuitive, at best.

Physicalists, however, can acknowledge that Mary has *some* sort of equal cognitive access to her experiences of unique red and red-31 in this situation, namely, she can discriminate the experience *tokens* in question and judge that they seem to be different. In this regard, Mary's experiences of unique red and red-31 are on a par. However, they can also point out that as soon as the roses are taken away, Mary (by hypothesis) will be left with the ability to remember and imagine unique red, but *not* red-31, and predict that, once this is realized, the intuition that she possesses equal knowledge of what it's like to have those experiences will fade.

It's true that if we take the ability to recognize or reidentify some property as necessary for having a (type) phenomenal concept of it, then these concepts constitute what may seem to be an odd amalgam of the coarse- and fine-grained. But once again, this result, though perhaps surprising, should not be damaging to physicalism. If we're to take psychology seriously in developing an account of phenomenal concepts—and of knowing what it's like—then these asymmetries are just the facts of life. And they should not be all that surprising, since there are many other familiar cases (smell, touch) in which the differences we can discriminate among various physical stimuli imperfectly reflect the differences that, from a physical point of view, there are.¹⁸

In addition, though the experiences we can discriminate in introspection will always outstrip the properties for which we can have type-demonstrative concepts, it's also a psychological fact that we can *increase* the range of these concepts, at least to some degree, through various sorts of instruction and practice. For example,

end p.96

people who have taken a course in music appreciation or wine tasting are able to recognize, or identify consistently, sounds or tastes that they were never able to recognize before, despite being able to discriminate them in simultaneous presentation. It seems

natural to describe what people get from these courses as increased knowledge of music or wine (or, focusing now on the correlative experiences, as increased first-person knowledge of various sounds or tastes). But if the ability to discriminate among items (or experiences) with different properties, when presented simultaneously, itself counts as having substantive knowledge, or cognitive grasp, of those properties, it's hard to describe what these people have learned.¹⁹ For this reason, too, the intuition that merely having, and being aware of, an experience provides substantive knowledge of the property experienced can be dispelled.

Antiphysicalists may suggest that they're better equipped to offer an account of how we can have substantive knowledge of what it's like to see red-31 (or first-person cognitive grasp of red-31), since the relation between phenomenal concepts and nonphysical properties can be tighter than the mere causal and dispositional relations between concepts and properties available to physicalists. If phenomenal properties are nonphysical, that is, they can be grasped in all their determinacy by a single act of looking in, and thus with one attentive look we can have full knowledge of them. And some physicalists may be tempted to match this suggestion by reconstructing a physicalistic version of whatever cognitive relation the dualists claim that we bear to our phenomenal states. As mentioned before, various theorists have suggested that we are "acquainted" with our phenomenal states, or that tokens of phenomenal properties themselves "partially constitute" our concepts of them, or that we possess a concept-forming mechanism that somehow "quotes" these property-tokens themselves when we think about them.

Physicalists, however, should resist these suggestions. First, they are metaphorical, and it's doubtful that any physicalistically acceptable version of them will satisfy dualists because the mechanisms invoked can be described in physicalistic language to the likes of Mary without giving her the phenomenal concepts in question. Second, they are unnecessary, since it's quite unclear that, on reflection, we would retain the conviction that we have substantive knowledge of what it's like to see red-31. As I've stressed already, there are many cases in which we can learn to recognize features of our experience that we couldn't reidentify before, and it seems natural to describe this as learning more about the properties in question. This belies the claim that we routinely acquire full knowledge of those properties at just one glance. And, as I've also stressed, any effects of our awareness of experiences produced by shades like the red-31 of a presented rose quickly fade when the rose is removed from sight, thus making it hard to hold that we know what it's like to see red-31 in any substantive sense. In short, on closer scrutiny, there is little intuitive support for the view that Mary has substantive knowledge of what it's like to see shades such as red-31.

However, there's a further question, namely, whether seeing colors for the first time gives Mary a richer mine of knowledge about color experience than can plausibly be explained by her acquisition of a new set of type-demonstrative concepts. Dualists, of course, have raised this question, but so have a number of physicalists. For example, in an earlier work (2002), I suggested that the only plausible way to explain Mary's acquisition of knowledge as the acquisition of new phenomenal concepts is to treat phenomenal concepts as hybrids with both recognitional and functional elements; that is, as relational descriptions of quality spaces with "slots" reserved for type-demonstratives that are normally acquired by having the experiences in question. On this view, in her black-and-

white room Mary learns that color experiences occupy certain places in a quality space, and when she steps out of the room and sees colors, she acquires the demonstratives to fill the slots. But because phenomenal concepts are in part relational, when Mary applies them in introspection to her own experiences, she is afforded knowledge of the rich relational network in which these (demonstrated) experiences are embedded—which is why it seems that she learns so much when she finally leaves her room.²⁰

It seems to me now, however, that a pure type-demonstrative view can give an analogous account of this phenomenon. To see this, let's examine what goes on when people come, through instruction and practice, to increase the range of their phenomenal concepts. For example, people who have taken a course in music appreciation or wine tasting will report that they are now able to recognize, or identify consistently, sounds or tastes that they were never able to recognize before.²¹ My aim, in the next section, is to ask what this training or instruction might involve, and what consequences it may have for our views about phenomenal concepts and “knowing what it's like.”

Increasing Our Stock of Phenomenal Concepts

As I see it, there are two (broadly characterizable) ways our recognitional abilities could be enhanced. The first involves the assimilation and application of explicit theoretical information about just where in our color quality space a certain target shade lies.²²

Suppose, for example, that when a number of colors are presented
end p.98

simultaneously, the person seeing them is taught to describe or think of red-31 as “the shade closer to unique red (along some dimension) than red-32”—and then perhaps as “the shade n jnd steps from unique red.” On this approach, an individual attempting to expand her range of color concepts can be seen as attempting to reidentify the experience of a certain target shade of red (red-31, say) *by consciously applying* information she has learned about its position in color quality space to her current experience. That is, she may eventually come to recognize a new instance of red-31 as the same shade she focused on before by imagining or remembering an experience of unique red (which is, by hypothesis, easy and natural for everyone), and comparing it to her current experience. (This is the sort of thing, I take it, that's often taught in wine-tasting courses: one is first given a wine with some easily identifiable taste, and then shown how to compare it to others that differ along various dimensions.) To do this, of course, she'd not only have to know that red-31 is, say, six jnd steps away from unique red, but also be able to remember or imagine the experience of unique red *and* the interval scale between unique red and red-31.²³

Now suppose that someone gets good at this procedure, and becomes able to reidentify the experience of red-31 consistently, without having to haul out this explicit comparative information or those imagined paradigms. Such things do happen, after all. Music appreciation and wine-tasting classes can, at least sometimes, provide people with the ability to home in, quickly and smoothly, on experiences they were incapable of reidentifying (and thus having type-demonstrative concepts of) before. In such cases, I suggest, the subject is deploying new type-demonstrative concepts.

One might suspect that the reidentifications one learns to make in this way depend on association or inference, and thus couldn't be epistemically “direct” enough to count as

the deployment of type-demonstrative concepts. But this need not be so. There are many familiar cases—distinctive kinds of vegetation, fabrics, styles of painting, or even unusual shapes—in which it takes time and instruction to acquire the ability to reidentify various phenomena, and thus the type-demonstrative concepts of those phenomena. In these cases, one might use association or inference to draw conclusions about the way things are from the way they seem: fabrics that feel like this are likely to be silk; paintings that look that way are likely to be Cezannes. But in acquiring the recognitional concepts themselves, one isn't using inference, but merely learning to attend to whatever distinctive features there are of the way things seem that qualify them as seeming *that way*.²⁴
end p.99

Can such concepts refer without introducing a new mode of presentation, given that they are, at least initially, informed by the explicit comparisons I described above? Yes, I suggest, under certain conditions. Concepts acquired as I've described might be taken to determine their referents in a way that is part demonstrative and part discursive; that is, they may be taken to refer to whatever property in fact bears the relevant relation to some paradigm—whether or not the subject using the concept would identify that property as the kind she had in mind. In this case, the concept wouldn't be a pure type-demonstrative concept (though it may include a demonstrative element). However, if the reference of the concept is determined solely by the subject's disposition to reidentify items as *another one of those*, then it should count as a pure type-demonstrative concept, regardless of how much explicit comparison, or other application of theory, was involved in its acquisition. In such a case the explicit information will be solely of heuristic value; it will shape a recognitional ability that by itself determines the referential reach of its associated (pure) type-demonstrative concept.²⁵

If theoretical information of the sort I've discussed figures in the acquisition and deployment of pure demonstrative concepts in this way, then physicalists can make the following claims. First, what someone gains from experience, in coming to know what it's like to have it, is—as the type-demonstrative theorists suggest—a new set of concepts distinct from any that she already possesses, and which denote directly, on the model of demonstratives, without needing to invoke metaphysically suspicious modes of presentation. And second, in at least *some* (and maybe most) cases, one has to know a lot about the properties in question, in particular their interrelations with others in the relevant quality space, to acquire the concepts—a phenomenon that accounts for the intuition that what one gains in knowing what it's like to see red or feel pain is interesting and substantial. So if this view of how we acquire new type-demonstrative concepts is plausible, the physicalist may be able to have it both ways: phenomenal concepts can have the referential role of type-demonstratives, but their acquisition may still require a person to know a lot about the interrelations among the properties they denote. But one may object that this view is *not* plausible in the least. What happens when we acquire new recognitional concepts is something much less cognitively complex, something that doesn't require anything like these explicit comparisons or other applications of theory that I've described. What goes on in wine-tasting or music appreciation classes—let alone what goes on when our recognitional

end p.100

dispositions are “naturally” shaped and enhanced by our interactions with the world—is closer to a “paradigm-foil” model of generalization learning than to anything I’ve described here.

That is, even in getting explicit instruction in how to reidentify a color or taste or sound, one is shown an example of the target item (and perhaps a foil that’s quite different along a certain dimension) and is told, when one tries to generalize to further instances, whether or not one did so correctly—without being told just *why* one’s attempt to generalize was correct or incorrect. In this case, one could increase one’s range of recognitional abilities, and thus phenomenal concepts, without appeal to any explicit comparisons or other theoretical information at all.

Maybe this is a more reasonable view of how recognitional abilities are enhanced by instruction and practice. And maybe it’s also a reasonable view of how we “naturally”—that is, without explicit instruction—come to recognize items in the world as belonging to one or another kind (and even, I suspect, of how innate proclivities to generalize may have been shaped by selective pressures).

If so, however, then the smooth deployment of type-demonstrative concepts would *still* require a significant amount of knowledge on the part of the subjects who deploy them. This wouldn’t be *explicit* knowledge of the salient features of the items recognized as “one of *that* kind,” or even of the similarities and differences between that kind and others, but rather the implicit knowledge of these similarities and differences that shaped (by the paradigm-foil method) the recognitional dispositions in question. Nonetheless, possession of these recognitional dispositions—and thus the (type-demonstrative) phenomenal concepts they determine—brings a lot to the table. In fact, it brings enough, I’d venture, to explain how Mary, in acquiring such dispositions after leaving her black-and-white room, manages to know so much.²⁶

end p.101

But even if the type-demonstrative view can account for the richness of Mary’s knowledge of what it’s like to see colors, there are other arguments against physicalism that need to be addressed, in particular, the “conceivability” arguments that depend on two-dimensional semantics, and the “distinct property argument,” which, if anything, has increased in subtlety since being posed to Smart by Max Black.²⁷

Zombies, Primary Intensions, and Modes of Presentation

Ever since Jackson (1993, 1998) and Chalmers (1996) introduced their versions of two-dimensional semantics,²⁸ and Chalmers (1996) used it to argue against physicalism, this framework and the argument it appears to support have become central to the debate.

Briefly, Chalmers takes the two-dimensional framework to provide a well-grounded explanation of why our ability to conceive of apparent counterexamples to identity statements (for example, to “H₂O = water”) does not show that these statements fail to express necessary truths. However, he argues, this explanation does not carry over to the case of phenomenal-physical identities. Thus, if we can conceive of zombies, we have no recourse but to conclude that they are possible, and thus that the “what it’s like” of

various experiences cannot be identified with physical properties.²⁹ I'll summarize both framework and argument as economically as possible, sketch the demonstrative theorists' response, and indicate some ways in which the focus of the debate has changed since Chalmers's initial presentation.

According to two-dimensional semantics, the meaning or intension of a term or concept is to be identified with a two-dimensional function (from possible worlds to denotations). The *primary* intension of a concept is a function that determines its extension in any world *w* (holding constant the way in which the reference of that concept is fixed in the actual world) if *w* is “considered as actual”—that is, if we assume that the way things are in *w* is the way things turned out to be in the actual world. The *secondary* intension of a concept, in contrast, is a function that determines the extension of the concept, in any world *w*, if *w* is “considered as counterfactual”—that is, if we assume that the actual world is just the way it is.³⁰

As should be clear, the primary and secondary intensions of natural kind concepts such as “water” diverge. In worlds “considered as actual,” a concept such as “water” will pick out whatever has the same essential properties as the stuff people in those worlds identify (by appeal to its qualitative properties) as “water.” Since these essential properties are microstructural, the primary intension of “water” assigns it H₂O here, XYZ on Twin Earth, and, in general, whatever meets these qualitative criteria in that world.³¹ But in “worlds considered as counterfactual,” “water” will always pick out H₂O, given that we understand it to denote (rigidly) just those substances that are microphysically similar to the stuff it denotes here. This explains why we can conceive of H₂O without water, even though it's a necessary truth that water is H₂O; what we're conceiving is a world, considered as actual, in which “water” picks out some substance which, though qualitatively like our water, is not H₂O.

But this explanation will not work, Chalmers (2003) argues, for what he calls our “pure” phenomenal concepts of experience. Just as with natural kind concepts, we understand the primary intensions of these pure phenomenal concepts to pick out, in each world, whatever has the same nature or essence as the experiences we identify by appeal to their qualitative characteristics. But in contrast with natural kind concepts, we take the nature or essence of these experiences to be those qualitative characteristics themselves. Thus, in each world, whether considered as actual or as counterfactual, phenomenal concepts pick out states qualitatively similar to the states we identify in these ways. That is, Chalmers argues, their primary and secondary intensions *coincide*.³² So, he concludes, my ability to conceive of a zombie can't be explained as my thinking that, in another world considered as actual, these phenomenal concepts pick out states qualitatively different from our

end p.103

own, and thus the conceivability of zombies must be taken as evidence of their possibility.³³

What Chalmers calls “pure” phenomenal concepts have an important role in this argument, and it's worth getting clear about what they are supposed to be. Chalmers gives

a “negative” characterization of them, as concepts distinct from those phenomenal concepts whose references are fixed by relations to external objects (e.g., “the kind of state produced by red things”) and also from demonstrative concepts whose references are fixed by acts of ostension (“the sort of experience I’m having now”) that could have picked out different items if the world had been different in various ways. He also gives a “positive” characterization of them, as concepts that “characterize ... the phenomenal quality as the phenomenal quality that it is” (2003, 226). What seems crucial for his argument, however—at least at first glance—is that these are concepts for which there is no room for variation in the values of their primary intensions, concepts whose primary and secondary intensions coincide.

Suppose, though, that the distinctive referential roles of Chalmers’s pure phenomenal concepts can be played by the special kinds of type-demonstratives that Loar (and others) take phenomenal concepts to be. Then physicalists could agree with Chalmers that the primary and secondary intensions of phenomenal concepts coincide, yet nonetheless invoke their irreducibly first-person or perspectival nature to explain why the conceivability of zombies does not threaten the identity of phenomenal and physical states or require there to be nonphysical modes of presentation of the physical states that phenomenal concepts denote. They would also have an explanation of why phenomenal-physical identities differ from other cases of inter-theoretic reduction and display an “explanatory gap”: namely, that only in the phenomenal-physical cases do the concepts in question have the first-personal, perspectival nature of demonstratives.

Our question, then, is whether type-demonstrative concepts that denote neural properties can play the distinctive referential role of pure phenomenal concepts. And it seems that the answer is “yes.” After all, these concepts are intended to pick out their target properties “directly,” without requiring modes of presentation that could “present” some *other* property in another world. Thus if my phenomenal type-demonstrative in fact denotes neural property *N*, it seems that it should pick out that same property in all (centered) worlds considered as actual, no matter what the rest of that world is like.³⁴

And if this is so, then there is no world in which something could *feel like this* and not instantiate property *N*—or vice versa.

end p.104

It may seem, however, that phenomenal demonstratives can’t possibly play the role of Chalmers’s pure phenomenal concepts. Recall his “positive” characterization of them as concepts that “characterize ... the phenomenal quality as the phenomenal quality that it is.”³⁵ It’s not obvious what this means, but, however it is further articulated, it’s clear that phenomenal type-demonstratives can’t meet this condition, given that they aren’t supposed to characterize phenomenal properties *as* anything at all.³⁶ Chalmers, later on in the same discussion (2003, 233), adds that, for pure phenomenal concepts, “the referent of the concept is somehow present inside the concept’s sense, in a way much stronger than in the usual cases of ‘direct reference.’” This, too, is far from clear. But however this condition is eventually spelled out, it’s unlikely that demonstrative theorists have the resources to avail themselves of it, since the “way” a referent is “present inside a concept’s sense,” by hypothesis, must be “much stronger” than the “usual cases” of direct

reference, in which reference is determined by the causal (and dispositional) relations the subject has to the item denoted.

Some physicalists, as already noted, have attempted to rise to the challenge by suggesting that phenomenal demonstratives must “quote,” or be “partially constituted by,” their referents,³⁷ but, once again, I believe this is futile. Physicalists, after all, have access only to materials such as causation, reliable correlation, and relations of physical inclusion or adjacency to reconstruct this notion of “presence,” and these, being “objective,” will prompt the same questions initially asked about Mary.³⁸ That is, dualists will surely object that Mary, in her black-and-white room, could know all these things about the relation of normal people's phenomenal concepts of red to the property of phenomenal red—and still not know what it's like to see red. Indeed. So why not stop the explanation earlier: if physicalists are convinced that treating phenomenal concepts as type-demonstratives can answer questions about Mary, the conceivability of zombies, and the explanatory gap, then they should reject the claim that phenomenal concepts require some sort of “presence” of, or “acquaintance” with, or “partial constitution” by the quality denoted, since this claim is backed only by the intuitions that they have already explained away.

There is a further question, however, that demonstrative theorists must address. Suppose, as I've claimed, that phenomenal type-demonstratives are concepts whose primary and secondary intensions coincide. But according to most theories of reference, so are scientific concepts such as “neural state *N*,” since these concepts, in worlds considered either as actual or counterfactual, seem capable of denoting

end p.105

only the neural states in question.³⁹ Now consider an identity statement such as “*this* (pointing in, under the appropriate conditions) = neural state *N*.” Such a statement should have a *necessary primary intension*; that is, it should be true in all worlds considered as actual. However, Chalmers (2006) argues explicitly that an identity statement *S* has a necessary primary intension if, and only if, it is a priori, and it's clear that a statement such as the one mentioned above is not a priori. This worry is analogous to the one directed to a stronger version of the distinct property argument, recently presented by White (1999, this volume, chap. 11) in which distinct properties are taken to be required for accounting not merely for how physical and phenomenal concepts could denote the same property, but for how people could *rationaly doubt* whether experiences are neural states.

Chalmers's argument, briefly, is that if *S* is not a priori, then it's conceivable—that is, epistemically possible—that *S* is false, and thus there is a possible world that verifies $\sim S$, and so *S* can't have a necessary primary intension. But now we are back to familiar territory: Physicalists who endorse the demonstrative account argue that in cases involving phenomenal concepts, a world that's conceivable (epistemically possible) need not be possible, and they point to the conceptual irreducibility of demonstratives as an explanation of why this is so. Antiphysicalists counter that a response like this makes a special exception for phenomenal-physical identities, since in all other cases of theoretical identity statements, the epistemic possibility of an entity that satisfies one term, but not the other, is evidence against the identity. And physicalists respond that it's

unfair to demand that a theory that acknowledges and explains the existence of a special “explanatory gap” in the case of phenomenal-physical identities should also be required to close it. Traversing this familiar territory, of course, may lead to a familiar impasse, but this, I’m afraid, goes with the territory traversed.⁴⁰
end p.106

There is a final worry about whether phenomenal concepts can be regarded as introspectively deployed type-demonstratives, namely, that it seems possible for me to use a type-demonstrative “*that* (kind of experience)” to ostend the same phenomenal property twice in quick succession—or even simultaneously—and yet have reason to doubt whether *that* is (the same property as) *that*. Hawthorne (chap. 10, this volume) presents a compelling example. Suppose a subject is told by scientists that the qualia on the right side of her visual field may “dance”—that is, shift from red to green—without her knowledge, during a certain time period. This subject, at least arguably, would thus have reason to doubt that *this* (pointing in to a reddish quale on the left) = *this* (simultaneously pointing in to a reddish quale on the right), even when she’s inclined to assert this and it’s true.⁴¹ This is reason, it may seem, for thinking that the subject must be using two distinct concepts—that is, concepts that denote via distinct modes of presentation, for otherwise there’s no explanation of how she could *rationaly doubt* that “*that* = *that*” is true.⁴²

Physicalists, however, can reply that not every case in which one can rationally doubt a true identity statement of the form “ $x = y$ ” is a case in which “ x ” and “ y ” denote their referents via distinct modes of presentation. In particular, a subject’s “rational doubt” in a scenario like Hawthorne’s can be shown to have different grounds and thus pose no threat to physicalism. Consider once again the scenario in question, in which a subject, attentively introspecting, twice deploys a type-demonstrative, but is told by an experimenter that her qualia may shift without her knowledge. And consider for a moment just how unusual such a situation would be. The subject, by hypothesis, is employing one type-demonstrative concept twice, which means that she is twice picking out a property which she is disposed to *reidentify*. In addition, she is deploying her concepts simultaneously, or in quick succession—and is paying attention. But she’s told by an authority that the qualia on one side of her visual field may shift without her knowledge. Given all this, her “rational doubt” can be attributed to the fact that her normal introspective confidence has been challenged by someone whom she thinks it rational to trust, and not to a difference in how the denoted property is “mediated” or “presented” on these occasions. That is, the subject could be wondering whether, despite her stable inclination to classify her experience-tokens as experiences of the same kind—that is, despite the fact that she *seems* to be using the same type-demonstrative concepts—she could nonetheless, for some arcane reason, be wrong, and thus be using *different* concepts.

The mistake that it’s possible for a rational subject to make, on this account, would be metalinguistic (or metaconceptual), and it may seem epistemically odd that introspecting subjects can be mistaken about whether they’re using the same concepts in their thoughts

about their own phenomenal states. But when concept difference and identity are determined “externally”—that is, by the features of what's denoted—this shouldn't be unexpected, even when the subject matter is one's own mental states.⁴³ This consequence may further motivate the attempt to make the relation between phenomenal concepts and denoted properties “closer” than possible on the pure demonstrative view, by requiring concepts to provide “acquaintance” with the properties they denote, or to “quote” them, or to be “partially constituted” by these properties. But, again, none of these strategies will help physicalists because no account of these relations that is consistent with physicalism can ensure that subjects, in thinking that their concepts pick out repeatable properties (as opposed, perhaps, to tokens, or instances, of phenomenal properties), must be correct. Physicalists, however, should be perfectly sanguine about this lack of infallibility, for, once again, it goes with the territory.

Does this entail that “*that* = *that*,” if true, is true a priori? No. As I've noted above, one doesn't have to claim that a statement is a priori to claim that its constituent terms converge in their primary intentions, and this is the only claim that the physicalist who is a demonstrative theorist needs.⁴⁴ In short, what physicalists need, and demonstrative theorists provide, is an account that individuates concepts by appeal to the properties demonstrated—that is, their contents. This provides whatever it takes to give them the same primary intension in all possible worlds, and for their primary and secondary intensions to coincide. They don't need any “closer” relation between concept and property, which, once again, is fortunate, since physicalistic versions of “constitution” or “acquaintance” are unlikely to do the job.

Thus, it seems as if the demonstrative view can account for the conceivability of zombies, the worries about distinct modes of presentation, and the richness of Mary's knowledge, and can narrow the explanatory gap while explaining why it can't be completely closed. Is it therefore a satisfactory view of phenomenal concepts? At the beginning of this chapter, I suggested that it may need reinforcement

end p.108

by a functional account to deal with what Block (2002) has called the “harder problem” of conscious experience. But this is (at least arguably) a harder problem—for which we can only begin to do the groundwork here.

Acknowledgments

I presented an early version of parts of this chapter at the University of California—Riverside in November 2001. I thank audiences there for helpful comments, and I am especially grateful to Eric Switzgebel. A later version was presented to the Chalmers/Hoy NEH seminar in July 2002. There I had a tremendously helpful discussion with many participants; I am particularly grateful to Dave Chalmers and Joseph Levine for their comments and suggestions. I'm also indebted to Brian Loar, who has discussed these issues with me over many years—and whose views I hope I'm not misrepresenting here. Finally, I want to thank Sven Walter and Torin Alter, whose suggestions and guidance led to many improvements in this chapter's substance and style.

References

- Austin, D. (1990). *What is the Meaning of 'This'?* Ithaca, N.Y.: Cornell University Press.
- Balog, K. (2002). The "Quotational Account" of Phenomenal Concepts. Unpublished.
- Bealer, G. (1987). Philosophical Limits of Scientific Essentialism. *Philosophical Perspectives* 1: 289–365. [Link ▶](#)
- Bealer, G. (2002). Modal Epistemology and the Rationalist Renaissance. In *Conceivability and Possibility*, ed. T. Gendler and J. Hawthorne: 71–126. Oxford: Oxford University Press.
- Block, N. (1996). Mental Paint and Mental Latex. *Philosophical Issues* 7: 1–17.
- Block, N. (2002). The Harder Problem of Consciousness. *Journal of Philosophy* 99, 1–35.
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Chalmers, D. J. (2003). The Content and Epistemology of Phenomenal Belief. In *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic: 220–72. Oxford: Oxford University Press.
- Chalmers, D. J. (2006). The Foundations of Two-Dimensional Semantics. In *Two-Dimensional Semantics: Foundations and Applications*, ed. M. Garcia-Carpintero and J. Macia. New York: Oxford University Press. An abridged version of this chapter, Epistemic Two-Dimensional Semantics, is in *Philosophical Studies* 118, 1–2, 2004: 153–226.
- Davies, M., and Humberstone, L. (1980). Two Notions of Necessity. *Philosophical Studies* 38: 1–30. [Link ▶](#)
- Hardin, C. (1988). *Color for Philosophers*. Indianapolis: Hackett.
- Horgan, T., and Graham, G. (2000). Mary, Mary, Quite Contrary. *Philosophical Studies* 99: 59–87. [Link ▶](#)
- Horgan, T., and Tienson, J. (2001). Deconstructing New-Wave Materialism. In *Physicalism and Its Discontents*, ed. C. Gillett and B. Loewer: 307–18. Cambridge: Cambridge University Press.
- Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36. [Link ▶](#)
- Jackson, F. (1993). Armchair Metaphysics. In *Philosophy in Mind*, ed. J. O'Leary-Hawthorne and M. Michael: 23–42. Dordrecht: Kluwer.
- end p.109
- Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.
- Kripke, S. (1972). Naming and Necessity. In *The Semantics of Natural Language*, ed. G. Harman and D. Davidson. Dordrecht: Reidel. Reprinted as *Naming and Necessity*. Cambridge: Harvard University Press, 1980.
- Levin, J. (1991). Analytic Functionalism and the Reduction of Phenomenal States. *Philosophical Studies* 61: 211–38. [Link ▶](#)

- Levin, J. (2002). Is Conceptual Analysis Needed for the Reduction of Qualitative States? *Philosophy and Phenomenological Research* 64: 571–91.
- Levine, J. (1983). Materialism and Qualia: the Explanatory Gap. *Pacific Philosophical Quarterly* 64: 354–61.
- Levine, J. (1993). On Leaving Out What It's Like. In *Consciousness: Philosophical and Psychological Essays*, ed. M. Davies and G. Humphreys: 121–36. Oxford: Blackwell.
- Levine, J. (1998). Conceivability and the Metaphysics of Mind. *Noûs* 32: 449–80.
- Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview. Revised version in *The Nature of Consciousness*, ed. by N. Block, O. Flanagan, and G. Güzeldere: 597–616. Cambridge: MIT Press, 1997.
- Loar, B. (2003). Qualia, Properties, Modalities. *Philosophical Issues* 13: 113–29. [Link](#)
- Lycan, W. (2003). Perspectival Representation and the Knowledge Argument. In *Consciousness: New Philosophical Essays*, ed. Q. Smith and A. Jokic: 384–95. Oxford: Oxford University Press.
- McLaughlin, B. (2001). In Defense of New Wave Materialism: A Response to Horgan and Tienson. In *Physicalism and Its Discontents*, ed. C. Gillett and B. Loewer: 319–30. Cambridge: Cambridge University Press.
- Nagel, T. (1974). What Is It Like to Be a Bat? *Philosophical Review* 83: 435–50. [Link](#)
- Nida-Rümelin, M. (1995). What Mary Couldn't Know: Belief about Phenomenal States. In *Conscious Experience*, ed. T. Metzinger: 219–41. Exeter: Imprint Academic.
- Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press. [Link](#) [OSO X-Reference](#)
- Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press.
- Raffman, D. (1995). On the Persistence of Phenomenology. In *Conscious Experience*, ed. T. Metzinger: 293–308. Exeter: Imprint Academic.
- Stalnaker, R. (1978). Assertion. *Syntax and Semantics* 9: 315–32.
- Stalnaker, R. (2003). Conceptual Truth and Metaphysical Necessity. In *Ways a World Might Be*: 201–15. Oxford: Oxford University Press. [Link](#) [OSO X-Reference](#)
- Sturgeon, S. (2000). *Matters of Mind*. London: Routledge.
- Tye, M. (1995). *Ten Problems of Consciousness*. Cambridge: MIT Press.
- Tye, M. (2000). *Color, Consciousness, and Content*. Cambridge: MIT Press.
- Tye, M. (2003). A Theory of Phenomenal Concepts. In *Minds and Persons*, ed. A. O'Hear: 91–105. Cambridge: Cambridge University Press. [Link](#)
- White, S. (1986). Curse of the Qualia. *Synthese* 68: 333–68. [Link](#)
- White, S. (1999). Why the Property Dualism Argument Won't Go Away. Unpublished. Available at <http://www.nyu.edu/gsas/dept/philo/courses/consciousness/papers/WHYPDAW.html>
- end p.110

seven Phenomenal and Perceptual Concepts

Papineau David

Phenomenal concepts are common coin among nearly all contemporary philosophers working on consciousness. They are recognized both by ontological dualists, who take them to refer to distinctive nonmaterial (phenomenal) properties, and by the majority of contemporary materialists, who respond that phenomenal concepts are distinctive only at a *conceptual* level and refer to nothing except material properties that can also be referred to using nonphenomenal material concepts.

In speaking of the majority of contemporary materialists, I have in mind the school of thought that David Chalmers (2003a) has dubbed “type-B physicalism.” In effect, type-B physicalism is a concession to the classic antimaterialist arguments of Frank Jackson (1986) and Saul Kripke (1972). Older (type-A) physicalists took all concepts of conscious states to be functional concepts—that is, concepts that referred by association with causal roles. Because of this, they denied the initial premises of Jackson's and Kripke's arguments. In response to Jackson's “Mary” argument, they argued that any functional concepts of conscious states would have been available to Mary before she left her room, so that there was no sense in which she acquired any new knowledge of “what it is like” to see something as red. In a related move, they responded to Kripke's argument by insisting that it was inconceivable, and thus obviously impossible, that a being could be fully physically identical to humans and yet lack consciousness. However, these responses to Jackson and Kripke are now widely agreed to be unsatisfactory. It seems clear that the preemergence Mary does lack some concepts of color experiences and, moreover, that zombies are at least conceivable. By recognizing phenomenal concepts, type-B physicalists aim to concede this much to Jackson and Kripke. At the same time, they argue that once we do recognize phenomenal concepts, then we can see that the subsequent stages of Jackson's and Kripke's arguments do not provide a valid route to ontologically dualist conclusions (cf. Loar 1990/97; Papineau 2002, chaps. 2 and 3). What is the nature of phenomenal concepts? Here there is far less consensus. Among those who trade in phenomenal concepts, some take them to be *sui generis*

end p.111

(Tye 2003, Chalmers 2003b), whereas others have variously likened them to recognitional concepts (Loar 1990/97), to demonstratives (Horgan 1984, Papineau 1993a, Perry 2001), or to quotational terms (Papineau 2002, Balog 2002).

In my *Thinking about Consciousness* (2002), I developed a “quotational-indexical” account of phenomenal concepts along roughly the following lines. To have a phenomenal concept of some experience, you must be able introspectively to focus on it when you have it, and to re-create it imaginatively at other times; given these abilities, you can then form terms with the structure *the experience*: —, in which the gap is filled either by a current experience or by an imaginative re-creation of an experience; these terms then constitute a distinctive way of referring to the experience at issue.

In that book, I argued that this account of phenomenal concepts not only allows a satisfactory materialist response to Jackson's and Kripke's arguments but also explains why dualism seems so compelling even to those unfamiliar with those arguments. According to my analysis, we all experience a basic “intuition of mind-brain distinctness,” which is prior to any philosophical investigation (and indeed which lends a

spurious plausibility to the standard antimaterialist arguments, by independently adding credibility to their conclusions). However, once we understand the structure of phenomenal concepts, I argued, we can see how this intuition arises and why it provides no real reason to doubt materialism.

In this chapter, I want to return to the topic of phenomenal concepts. It now seems to me that the treatment in *Thinking about Consciousness* was inadequate in various respects. Here I want to try to improve on that account. In particular, I shall develop an extended comparison of phenomenal concepts with what I shall call “perceptual concepts,” hoping thereby to throw the nature of phenomenal concepts into clearer focus.

Though the position I develop in this chapter will involve some significant revisions of the claims made in my book, I think that the main arguments in the book are robust with respect to these revisions. In particular, the responses to Jackson and Kripke stand pretty much as before, and an explanation of the persistent “intuition of distinctness” continues to be available.

The revised account also enables me to deal with a common worry about phenomenal concepts.¹ Suppose Mary has come out of her room, seen a red rose, and as a result acquired a phenomenal concept of the experience of seeing something red (though she may not yet know that this experience is conventionally so-called). On most accounts of phenomenal concepts, including the one developed in my book, any exercise of this phenomenal concept will demand the presence of the experience itself or an imaginatively re-created exemplar thereof. The trouble, however, is that it seems quite possible for Mary to think truly, using her new phenomenal concept, *I am not now having that experience (nor re-creating it in my imagination)*—but this would be ruled out if any exercise of her phenomenal concept did indeed depend on the presence of the experience or its imaginative re-creation. The revised account of phenomenal concepts developed here does not require this, and so can explain Mary's problematic thought.

The rest of this chapter contains four sections. The next two analyze perceptual and phenomenal concepts, respectively. The penultimate section checks that my revised account of phenomenal concepts will still serve to block the standard arguments for dualism. The final section defends my position against a recent argument by David Chalmers against the whole type-B strategy of defending physicalism by appeal to phenomenal concepts.

Perceptual Concepts

Perceptual Concepts Are Not Demonstrative

Let me turn away from phenomenal concepts for a while and instead consider perceptual concepts. Getting clear about perceptual concepts will stand us in good stead when we turn to the closely related category of phenomenal concepts.

We can start with this kind of case. You see a bird at the bottom of your garden. You look at it closely, and at the same time think, I haven't seen *that* in here before. Later on, you can recall the bird in visual imagination, perhaps thinking, I wonder if *that* was a

migrant. In addition, on further perceptual encounters with birds, you sometimes take some bird to be the same bird again, and you can again form further thoughts about it, such as *that* bird has a pleasant song. (Let me leave open for the moment whether you are thinking of a particular bird or a type of bird; I shall return to this shortly.)

In examples of this sort, I shall say that subjects are exercising *perceptual concepts*.

Perceptual concepts allow subjects to think about perceptible entities. Such concepts are formed when subjects initially perceive the relevant entities, and they are reactivated by later perceptual encounters. Subjects can also use these concepts to think imaginatively about those entities even when the entities are not present.

It is tempting to view concepts of this kind as “demonstrative.” For one thing, it is natural to express these concepts using demonstrative words, as the above examples show (e.g., “... *that* ...”). Moreover, uses of perceptual concepts involve a kind of perceptual attention or imaginative focus, and this can seem analogous to the overt pointing or other indicative acts that accompany the use of verbal demonstratives.

However, I think it is quite wrong to classify perceptual concepts as demonstratives. If anything is definitive of demonstrative terms, it is surely that they display some species of characterlikeness. By this I mean that the referential value of the term depends on context: the selfsame term will refer to different items in different contexts. However, there seems nothing characterlike about the kind of perceptual concept illustrated in the above examples. Whenever it is exercised, your perceptual concept refers to the *same* bird. When you use the concept in question, you don't refer to one bird on the first encounter, yet some possibly different bird

end p.113

when later encountering or visually imagining it. Your concept picks out the same bird whenever it is exercised.

It is possible to be distracted from this basic point by failing to distinguish clearly between perceptual concepts and their linguistic expression. If I want to express some perceptual thought in language, then there may be no alternative to the use of demonstrative *words*. In order to convey my thought to you, I may well say, “*That* bird has a pleasant song,” while indicating some nearby bird. And I agree that the words here used—“that bird”—are demonstrative, in that they will refer to different birds in different contexts of use. But this does not mean that my concept itself is demonstrative. As I have just urged, my concept itself will refer to the same bird whenever it is exercised.

The reason we often resort to demonstrative words to convey thoughts involving nondemonstrative perceptual concepts is simply that there is often no publicly established linguistic term to express our concept. In such cases, we can nevertheless often get our ideas across by demonstratively indicating some instance of what we are thinking about. Of course, this possibility assumes that some such instance is available to be demonstrated—if there isn't, then we may simply find ourselves unable to express what we are thinking to an audience.

By insisting that perceptual concepts are not demonstrative, even if the words used to express them are, I do not necessarily want to exclude characterlikeness from every aspect of the mental realm. Millikan (1990) has argued that mental indexicality plays no ineliminable role in the explanation of action, against Perry (1979) and much current

orthodoxy, and I find her case on this particular point persuasive. Even so, I am open to the possibility that primitive mental demonstratives may play some role in preconceptual attention (e.g., what was *that*?) and also to the possibility that there may be characterlike mental terms constructed with the help of predicates (e.g., I'm frightened of *that dog*—meaning the dog in the corner of the room).² In both these kinds of case I allow that the italicized expressions may express genuinely characterlike mental terms—that is, repeatable mental terms that have different referents on different occasions of use. My claim in this section has only been that perceptual concepts in particular are not characterlike in this sense but carry the same referent with them from one occasion of use to another.

Perceptual Concepts as Stored Templates

I take perceptual concepts to involve a phylogenetically old mode of thought that is common to both humans and animals. We can helpfully think of perceptual concepts as involving stored *sensory templates*. These templates will be set up on initial encounters with the relevant referents. They will then be reactivated on later perceptual encounters, via matches between incoming stimuli and stored templates—
end p.114

perhaps the incoming stimuli can be thought of as “resonating” with the stored patterns and thereby being amplified. Such stored templates can also be activated autonomously even in the absence of any such incoming stimuli—these will then constitute “imaginative” exercises of perceptual concepts.³

The function of the templates is to accumulate information about the relevant referents and thereby guide the subject's future interactions with them. We can suppose that various items of information about the referent will become attached to the template as a result of the subject's experience. When the perceptual concept is activated, these items of information will be activated, too. They may include features of the referent displayed in previous encounters. Or they may simply comprise behavioral information in the form of practical knowledge that certain responses are appropriate to the presence of the referent. When the referent is reencountered, the subject will thus not only perceive it as presently located at a certain position in egocentric space, but will also take it to possess certain features that were manifested in previous encounters, but may not yet be manifested in the reencounter. Imaginative exercises of perceptual concepts may further allow subjects to process information about the referent even when it is not present.

Note how this function of carrying information from one use to another highlights the distinction between perceptual concepts and demonstratives. Demonstrative terms do not so carry a body of information with them, for the obvious reason that they refer to different entities on different occasions of use. Information about an entity referred to by a demonstrative on one occasion will not in general apply to whatever entity happens to be the referent the next time the demonstrative is used. By contrast, perceptual concepts

are suited to serve as repositories of information precisely because they refer to the same thing whenever they are exercised.

Perceptual Semantics

I have said that perceptual concepts refer to perceptible entities. However, what exactly determines this relation between perceptual concepts, conceived as stored sensory templates, and their referents? In particular, what determines whether such a concept refers to a type or a token? I suggested earlier that you might look at a bird, form some stored sensory template, and then use it to think *either* about that particular bird *or* about its species. But what decides between these two referents? At first pass, it seems that just the same sensory template might be pressed into either service.

Some philosophers think of perceptual concepts as “recognitional concepts” (Loar 1990/97). This terminology suggests that perceptual concepts should be viewed as referring to whichever entities their possessors would recognize as satisfying them. A stored sensory template will refer to the entity that will activate it when encountered. If none but some particular bird will activate some template, then that
end p.115

particular bird is the referent. If any member of a bird species will activate a template, then the species is the referent.

This recognitional account would serve adequately for most of the further purposes of this chapter. But in fact it is a highly unsatisfactory account of perceptual reference. Now that I have raised this topic, I would like to digress briefly and explain how we can do better.

First, let me briefly point out the flaws in the recognitional account. For a start, it's not clear that recognitional abilities are fine-grained enough to make the referential distinctions we want. Could not two people have just the same sensory template and so be disposed to recognize just the same instances, and yet one be thinking about a particular bird, and the other about the species? It is not obvious, to say the least, that my inability to discriminate perceptually between the bird in my garden and its conspecifics means that I must be thinking about the whole species rather than my particular bird. Nor, conversely, is it obvious that I must be thinking of my bird rather than its species if I mistakenly take some idiosyncratic marking of my bird to be a characteristic of the species. In any case, the equation of referential value with recognitional range faces the familiar problem that it seems to exclude any possibility of *misrecognition*: if the referent of my perceptual concept is that entity which includes all the items I recognize as satisfying the concept, then there is no room left for me to misapply the concept perceptually. However, this isn't what we want—far from guaranteeing infallibility, possession of perceptual concepts seems consistent with very limited recognitional abilities.

I think we will do better to approach reference by focusing on the *function* of perceptual concepts rather than their actual use. As I explained in the last subsection, the point of

perceptual concepts is to accumulate information about certain entities and make it available for future encounters. Given this, we can think of the referential value of a perceptual concept as that entity which the perceptual concept has the function of accumulating information about. Give or take a bit, this will depend on two factors: the *origin* of the perceptual concept, and the *kind* of information that gets attached to it. Let me take the second factor first. Note that the kind of information that it is appropriate to carry from one encounter to another will vary, depending on what sort of entity is at issue.⁴ For example, if I see that some bird has a missing claw, then I should expect this to hold on other encounters with that particular bird, but not across encounters with other members of that species. By contrast, if I see that a bird eats seeds, then this information is appropriately carried over to other members of the species. The point is that different sorts of information are projectible across encounters with different types of entity. If you are thinking about some metal, you can project melting point from one sample to another, but you can't do the same thing with the shape of the samples. If you are thinking about some species of shellfish, you can project shape but not size. If you are thinking about individual humans, you can project ability to speak French, but not shirt color. And so on.

end p.116

Given this, we can think of the referents of perceptual concepts as determined *inter alia* by what *sort* of information the subject is disposed to attach to those concepts. If the subject is disposed to attach particular-bird-appropriate information, then the concept refers to a particular bird, whereas if the subject is disposed to attach bird-species-appropriate information, then reference is to a species. In general, we can suppose that the concept refers to an instance of that kind to which the sort of information accumulated is appropriate.

To make this suggestion more graphic, we might think of the templates corresponding to perceptual concepts as being manufactured with a range of "slots" ready to be filled by certain items of information. Thus a particular-bird-concept will have slots for bodily injuries and other visible abnormalities; a particular-person-concept will have slots for languages spoken; a metal-concept will have a slot for melting point; and so on. Which slots are present will then determine which kind of entity is at issue.

The actual referent will then generally be whichever instance of that kind was responsible for originating the perceptual concept. As a rule, we can suppose that the purpose of any perceptual concept is to accumulate information about the item (of the relevant kind) that was responsible for its formation. This explains why there is a gap between referential value and recognitional range. I may not be particularly good at recognizing some entity. But if that entity is the source of my concept, then the concept's function is still to accumulate information about it.

Of course, if some perceptual concept comes to be regularly and systematically triggered by some entity other than its original source, and, as a result, information derived from this new entity comes to eclipse information about the original source, then no doubt the concept should come to be counted as referring to the new entity rather than the original source. But this special case does not undermine the point that a perceptual concept will normally refer to its origin, rather than to whichever entities we happen to recognize as fitting it.

Now that I have explained how it is possible for perceptual concepts to refer differentially to both particular tokens and general types, some readers may be wondering how things will work with subjects who have perceptual concepts both for some token and its type—for example, suppose that I have a perceptual concept both for some particular parrot and for its species. To deal with cases like this, we need to think of perceptual concepts as forming structured hierarchies. When someone has perceptual concepts both for a token and its type, the former will add perceptual detail to the latter, so to speak. The same will also apply when subjects have concepts of some determinate (*mallard*, say) of some determinable type (*duck*). In line with this, when some more detailed perceptual concept is activated, then so will any more general perceptual concepts that cover it, but not vice versa. Since any items of information that attach to the more general concepts will also apply to the more specific instances, this will work as it should, giving us any generic information about the case at hand along with any case-specific information.

Before proceeding, let me make it clear how I am thinking about the relationship between perceptual concepts and *conscious* perceptual experience. I want to equate conscious perceptual experiences with the activation of perceptual concepts, ascribable either to exogenous stimulation or to endogenous imagination. This does not necessarily mean that *all* activations of perceptual concepts are conscious. There may be states that fit the specifications of perceptual concepts given so far, but whose activations are too low-level to constitute conscious states—early stages of visual processing, say. My assumption will be only that there is some range of perceptual concepts whose activations constitute conscious perceptual experience.⁵ In line with this, I shall restrict the term “perceptual experience” to these cases—that is, I shall use “experience” in a way that implies consciousness. In addition, I shall also assume that the phenomenology of these states goes with the sensory templates involved, independently of what information the subject attaches to those templates or is or is not disposed to attach to them. (So if you and I use the same sensory pattern to think about a particular bird and a bird species, respectively, the what-it's-likeness of the resulting experiences will nevertheless be the same.)⁶

Perceptually Derived Concepts

The discussion so far has assumed that thoughts involving perceptual concepts will require that the subject actually be perceiving or imagining. In order for the perceptual concept to be deployed, the relevant stored template needs either to be activated by a match with incoming stimuli or to be autonomously activated in imagination.⁷

However, now consider this kind of case. You have previously visually encountered some entity—a particular bird, let us suppose—and have formed a perceptual concept of that bird. As before, you exercise this perceptual concept when you perceive something as the same bird again, or when you imagine the bird. However, now suppose that you think about the bird when it is not present, and without imaginatively re-creating your earlier perception. You simply think, “That bird must nest near here,” say, without any accompanying perceptual or imaginative act. I take it that such thoughts are possible.⁸ Having earlier established perceptual contact with some entity, you can subsequently refer to it without the active help

end p.118

of either perception or imagination. I shall say that such references are made via *perceptually derived concepts*.

Here is one way to think about this. Initially your information about some referent was attached to a sensory template. But now you have further created some nonperceptual “file” in which your store of information about that entity is also housed. This enables you to think about the entity even when you are not perceiving or imagining it. When you later activate the file, you automatically refer to the same entity as was referred to when the file was created.⁹

Perhaps the ability to create such nonperceptual files is peculiar to linguistic creatures. This is not to say that any such file must correspond to a term in a public language: you can think nonperceptually about things for which you have no name—for example, you may have no name for the bird that you think nests nearby. Still, it seems likely that the ability to think nonperceptually evolved with the emergence of language. In this connection, note that an ability to think about things that you have not perceived, and so cannot perceptually recognize or imagine, must play an essential part in mastery of a public language. For public languages are above all mechanisms that allow those who have firsthand acquaintance with certain items of information to share that information with others, which means that those who receive such information will often need to create nonperceptual “files” for entities they have never perceived. By contrast, languageless creatures will have no channels through which to acquire information about items beyond their perceptual ambit, and thus will have no need to represent those items nonperceptually. This provides reason to suppose that the ability to create non-perceptual “files” arrives only with the emergence of language. If this is right, then only language-using human beings will be able to transcend perceptual concepts proper by constructing what I am calling “perceptually derived concepts.” Of course, as noted at the beginning of this paragraph, humans will also sometimes use this ability to create nonperceptual “files” that correspond to no word in a public language. But when they do so, they may well be drawing on an ability that evolved with linguistic capacities.

Perhaps there is an issue about counting concepts here. I have distinguished between “perceptual concepts” and “perceptually derived concepts.” Do I therefore
end p.119

want to say that a thinker who has constructed a “perceptually derived concept” from a prior “perceptual concept” now has *two* concepts that refer to the same thing? From some perspectives, this might seem like double counting. In particular, it is not clear that the standard Fregean criterion of cognitive significance will tell us that there are two concepts here. After all, if the creation of a “perceptually derived concept” is simply a matter of housing your store of information in a nonperceptual file, and if any subsequently acquired information about the relevant referent automatically gets attached to both a sensory template and nonperceptual tag, then it seems that the subject will always make exactly the same judgments whether using the “perceptual” or “perceptually derived” concept, and so fail the Frege test for possession of distinct concepts. And this

would suggest that we simply have *one* concept here, not two, albeit a concept that can be exercised in two ways—perceptually and nonperceptually.¹⁰

There is no substantial issue here. To the extent that the flow of information between the two ways of thinking is smooth, the Frege test gives us reason to say that there is only one concept. On the other hand, to the extent that there are cognitive operations that distinguish a perceptually derived concept from its originating perceptual concept, there is a rationale for speaking of two concepts, and I shall do so when this is convenient.

Phenomenal Concepts

The Quotational-Indexical Model

Let me now turn to phenomenal concepts. Recall that my earlier quotational-indexical model viewed phenomenal concepts as having the structure *the experience*:—, where the gap was filled either an actual perceptual experience or an imaginative re-creation thereof. It now seems to me that this model ran together a good idea with a bad one. The good idea was to relate phenomenal concepts to perceptual concepts. The bad idea was to think that phenomenal concepts, along with perceptual ones, are some kind of “demonstrative.”

Let me first explain the bad idea. Suppose that perceptual concepts were demonstrative, contrary to my argument above. Then presumably they would be constructions that, on each occasion of use, referred to whichever item in the external environment was somehow salient to the subject. By analogy, if phenomenal concepts worked similarly, then they too would refer to salient items, but in the “internal” conscious environment. This led me to the idea that phenomenal concepts were somehow akin to the mixed demonstrative construction *that experience*. On this model, phenomenal concepts would employ the same general demonstrative construction (*that*) as is employed by ordinary mixed demonstratives, but the qualifier *experience* would function to direct reference inward, so to speak, ensuring that some salient element in the conscious realm is picked out. The “quotational” suggestion then depended on the fact that this demonstrated experience would itself

end p.120

be present in the realm of conscious thought, unlike the nonmental items referred to by most demonstratives. This made it seem natural to view phenomenal concepts as “quoting” their referents, rather than simply referring to distal items. Linguistic quotation marks, after all, are a species of demonstrative construction: a use of quotation marks will refer to *that word*, whatever it is, that happens to be made salient by being placed within the quotation marks. Similarly, I thought, phenomenal concepts can usefully be thought of as referring to *that experience*, whatever it is, that is currently made salient in thought.

However, this now seems to me all wrong. Not only is it motivated by a mistaken view of perceptual concepts, but it runs into awkward objections about the nature of the notion of *experience* used to form the putative construction *that experience*.

There seem two possible models for the concept of *experience* employed here. It might be *abstracted* from more specific phenomenal concepts (e.g., seeing something red, smelling roses, and so on); alternatively, it could be some kind of *theoretical* concept, constituted by its role in some theory of experiences. However, neither option seems acceptable.

The obvious objection to the abstraction strategy is that it presupposes such specific phenomenal concepts as seeing something red, smelling roses, and so on, when it is supposed to explain them. If we are to acquire a generic concept of experience via first thinking phenomenally about more specific experiences, and then abstracting a concept of what they have in common, then it must be possible to think phenomenally about the more specific experiences prior to developing the generic concept. But if thinking phenomenally about the more specific experiences requires us already to have the generic concept, as in the demonstrative account of phenomenal concepts, then we are caught in a circle.

What if our notion of experience is constituted by its role in some theory of experiences (our folk psychological theory perhaps)? Given such a theoretically defined generic concept of experience, there would be no barrier to then combining it with a general-purpose “that” to form demonstrative concepts of specific experiences. Since the generic concept wouldn't be derived by abstraction from prior phenomenal concepts of specific experiences, there would be no circle in using it to form such specific phenomenal concepts.

This picture may be cogent in principle, but it seems to be belied by the nature of our actual phenomenal concepts. If a generic concept of “experience” were drawn from something like folk psychological theory, then we could expect it to involve some commitment to the assumption that experiences are internal causes of behavior. Folk psychology surely conceives of experiences *inter alia* as internal states with characteristic causes and behavioral effects. But then it would seem to follow that anything demonstrated as *that experience*, where *experience* is the folk psychological concept, must analytically have some behavioral effects. Which specific behavioral effects *that experience* has needn't be analytic—you could know that all experiences have characteristic effects without knowing what specific effects *that experience* has—but still, it would be analytic that *that experience* had *some* behavioral effects. However, this doesn't seem the right thing to say about phenomenal concepts. There is surely nothing immediately contradictory in the idea that an experience picked out by some phenomenal concept has no subsequent effects on

end p.121

behavior or anything else. Epiphenomenalism about phenomenal states doesn't seem to be a priori contradictory.¹¹ Yet it would be, if our ways of referring to phenomenal states analytically implied that they had behavioral effects.

Phenomenal Concepts as Perceptual Concepts

I said above that my old model of phenomenal concepts ran together the good idea that phenomenal concepts are related to perceptual concepts with the bad idea that both kinds of concepts are “demonstratives.” Let me now try to develop the good idea unencumbered by the bad one.

My current view is that phenomenal concepts are simply special cases of perceptual concepts. Consider once more the example where I perceptually identify some bird and make a judgment about it (e.g., *That* is a migrant). I earlier explained how the perceptual concept employed here could either be a concept of an individual bird or the concept of a species. I want now to suggest that we think of phenomenal concepts as simply a further deployment of the same sensory templates, but in this case being used to think about perceptual experiences themselves rather than about the objects of those experiences. I see a bird, or visually imagine a bird, but now I think not about that bird or a species but about the experience, the conscious awareness of a bird.¹²

The obvious question is, what makes it the case that I am here thinking about an experience rather than about an individual bird or a species? However, we can give the same answer here as before. I earlier explained how the subject's dispositions to carry information from one encounter to another can decide whether a given sensory template is referring to an individual rather than a species, or vice versa: if the subject projects species-appropriate information, reference is to a species, whereas if the subject projects individual-appropriate information, reference is to an individual. So let us apply the same idea once more—if the subject is disposed to project *experience-appropriate* information from one encounter to another, then the sensory template in question is being used to think about an experience. For example, suppose I am disposed to project, from one encounter to another, such facts as that what I am encountering ceases when I close my eyes, goes fuzzy when I am tired, will be more detailed if I go closer, and so on. If this is how I am using the template as a repository of information, then I will be referring to the visual experience of seeing the bird rather than to the bird itself. More generally, if they are used in this kind of way—to gather experience-appropriate information, so to speak—the same sensory templates that are normally used to think about perceptible things will refer to experiences themselves.

Can phenomenal concepts pick out experiential particulars as well as types? In the perceptual case, as we have seen, there is room for such differential reference to both particular objects and to types, thanks to the possibility of differing dispositions carrying information from one encounter to another. In principle it may seem that the same sort of thing could work in the phenomenal case. The trouble, however, is that particular experiences, unlike ordinary spatiotemporal particulars, do not seem to persist over time in the way required for reencounters to be possible. Can the same *particular* pain, or *particular* visual sensation, or *particular* feeling of lassitude reoccur after ceasing to be phenomenally present? It is true that we often say things like “Oh dear, there's that pain again—I thought I was rid of it.” But nothing demands that we read such remarks as about quantitative rather than qualitative identity. Nothing forces us to understand them as saying that the same particular experience has reemerged, as it were, rather than that the same experiential type has been reinstantiated (note in particular that experiences do

not seem to allow anything analogous to the spatiotemporal tracking of ordinary physical objects). In line with this, note that information about experiences, as opposed to information about spatiotemporal particulars, does not seem to divide into items that are projectible across encounters with a particular and items that are projectible across encounters with a type.

Given all this, I am inclined to say that phenomenal concepts cannot refer differentially both to particulars and to types. Rather, they always refer to types—that is, to the kind of mental item that can clearly reoccur. As I am conceiving of perceptual and phenomenal concepts, the function of a concept is to carry information about its referent from one encounter to another—and it seems that only phenomenal types and not particulars can be reencountered.

The corollary is that when we do refer to particular experiences, we cannot be using our basic apparatus of phenomenal concepts, given that these are only capable of referring to phenomenal types. Rather, we must be invoking more sophisticated conceptual powers, such as the ability to refer by description (thus *the particular pain I am having now*, or *the particular experience of crimson I enjoyed at last night's sunset*).

Phenomenal Use and Mention

This model of phenomenal concepts as a species of perceptual concept retains one crucial feature from my earlier quotational-indexical model, namely, that phenomenal references to an experience will deploy an instance of that experience, and in this sense will *use* that experience in order to mention it.

To see why, think about what happens when a phenomenal concept is exercised. Some sensory template is activated and is used to think about an experience. This sensory activation will result from externally generated sensory stimuli or autonomous imaginative activity. That is, you will either be perceiving the environment
end p.123

or employing perceptual imagination; for example, either you will be perceiving a bird, or you will be perceptually imagining one. Except, when phenomenal thought is involved, this template is also used to think about perceptual experience itself and not only about the objects of perceptions. You look at a bird or visually imagine that bird, but now you use the sensory state to think about the *visual experience of seeing the bird*, and not only about the bird itself.

This means that any exercise of a phenomenal concept to think about a perceptual experience will inevitably involve either that experience itself or an imaginary re-creation of that experience. If we count imaginary re-creations as “versions” of the experience being imagined, then we can say that phenomenal thinking about a given experience will always *use* a version of that experience in order to *mention* that experience.

Note how this model accounts for the oft-remarked “transparency of experience” (Harman 1990). If we try to focus our minds on the nature of our conscious experiences, all that happens is that we focus harder on the objects of those experiences. I try to concentrate on my visual experience of the bird, but all that happens is that I look harder at the bird itself. Now, there is much debate about exactly what this implies for the nature of conscious experience (cf. Stoljar forthcoming). But we can bypass this debate here and simply attend to the basic phenomenon, which I take to be the *phenomenological*

equivalence of (a) thinking phenomenally *about* an experience, and (b) thinking perceptually *with* that experience. What it's like to focus phenomenally on your visual experience of the bird is no different from what it's like to see the bird.

On my model of phenomenal thinking, this is just what we should expect. I said at the end of the last section that the phenomenology of perceptual experiences is determined by which sensory template they involve, and not by what information they carry with them. I have now argued that just the same sensory templates underlie both perceptual experiences and phenomenal thoughts about those experiences. It follows that perceptual experiences and phenomenal thoughts about them will have just the same phenomenology. This explains why thinking phenomenally about your visual experience of a bird feels no different from thinking perceptually about the bird itself.

A Surprising Implication

The story I have told so far has an implication that some may find surprising. On my account, the semantic powers of phenomenal concepts would seem to depend on their cognitive function rather than on their phenomenal nature. I have argued that phenomenal concepts refer to conscious experiences because it is their purpose to accumulate information about those experiences. As it happens, exercises of such concepts will in part be constituted by versions of the conscious experiences they refer to, and so will share the what-it's-likeness of those experiences. But this latter, phenomenal fact seems to play no essential role in the semantic workings of phenomenal concepts. To see this, suppose that we had evolved to attach information about conscious experiences to states other than sensory templates—to words in some language of thought, perhaps. Wouldn't these states refer equally to

end p.124

experiences, and for just the same reason, even though their activation did not share the phenomenology of their referents? However, this may seem in tension with the idea that phenomenal concepts involve some distinctive mode of phenomenal self-reference to experiences. If the phenomenality of phenomenal concepts is incidental to their referential powers, then in what sense are they distinctively phenomenal? (Cf. Block, chap. 12, this volume.)

Note that my earlier quotational-indexical account of phenomenal concepts is not open to this kind of worry. On that account, phenomenal concepts used experiences as *exemplars* rather than as ways of implementing a cognitive role. Given this, it is essential to the phenomenal concept of *seeing something red*, say, that “quotes” some version of that experience, just as it is essential to the quotational referring expression “ ‘zymurgy’ ” that it contain the last word in the English dictionary within its single quotation marks. Thus on the quotational-indexical account, there is no question of some state referring to an experience in the same way that a phenomenal concept does without containing that experience.

Note also that the worrisome implication is not peculiar to the particular theory of the semantics of phenomenal concepts that I have defended in this chapter. It will arise on any theory that makes the semantic powers of phenomenal concepts a matter of their conceptual role, or their informational links to the external world, or any other facet of their causal-historical workings. For any theory of this kind will make it incidental to the

referential powers of phenomenal concepts that they have the same phenomenology as their referents. Any such theory leaves it open that some other state, with a different or no phenomenology, could have the same causal-historical features and thus refer to experiences for the same reason that phenomenal concepts do.

My response to this worry is that there is no real problem here. On my account, phenomenal concepts do indeed refer because of their cognitive function, not because of their phenomenology, and therefore other states with a different or no phenomenology, but with the same cognitive function, would refer to the same experiences for the same reasons. I see nothing wrong with this. Of course, it is a further question whether we would wish to include any such nonphenomenological states within the category of “phenomenal” concepts, given their lack of what-it's-likeness (cf. Tye 2003). But this is no ground for denying that they would refer to experiences for just the same reason that phenomenal concepts do.

I shall come back to the issue of what counts as a “phenomenal” concept in the next section. But first let me ask a somewhat different question. Given that other things could in principle play the cognitive role that determines reference to experiences, *why* do we use experiences themselves for this purpose? What is it about conscious experiences that makes them such a good vehicle for referring to themselves?

One possible answer is that this use of experiences is somehow well suited to answering certain questions. To adapt an example of Michael Tye's (2003: 102), suppose that we are wondering whether the England one-day cricket strip is visually darker than the Indian one. By thinking phenomenally about these colors, we will generate versions of the relevant experiences and be in a position to compare them directly.

end p.125

This makes some sense, but I think a simpler answer may be possible. Consider the analogous question: why do we use *perceptual* experience to represent perceptible items such as people, physical objects, animals, plants, shapes, colors, and so on? After all, in this case too the referential powers of these states are presumably determined by some type of cognitive role, which could in principle have been played by something other than perceptual experiences themselves. Here the obvious answer seems to be that perceptions are especially good for thinking about perceptible entities simply because they are characteristically activated by those entities, and so are well suited to feature in judgments that those entities are present. It would unnecessarily duplicate cognitive mechanisms to use the perceptual system to identify perceptible entities and yet use something other than perceptual experiences as the vehicle for occurrent thoughts that imply that those entities are present.

This idea applies all the more in the phenomenal case. Conscious experiences are excellent vehicles for thinking about those selfsame experiences simply because they are automatically present whenever their referents are. The fact that we use experiences to think about themselves means that we don't have to find other cognitive resources to frame occurrent thoughts about the presence of experiences.

The Knowledge Argument

In the last subsection, I raised the question of what exactly qualifies a concept as “phenomenal.” I have no definite answer to this essentially definitional question. Far more important, from my point of view, is whether phenomenal concepts as introduced so far provide effective answers to the standard antiphysicalist arguments. I shall now aim to show that they do this. In the course of doing so, however, I shall highlight those features of phenomenal concepts that are important to their serving this philosophical function. I’ll leave open which features of phenomenal concepts are essential to their counting as “phenomenal.” Rather, I’ll focus on which features matter to the philosophical arguments.

Let me begin with Frank Jackson’s knowledge argument. Here the type-B physicalist response is that there is indeed a sense in which Mary doesn’t “know what seeing red is like” before she comes out of her room, despite her voluminous material knowledge. But this is not because there is any objective feature of reality that her material knowledge fails to capture. Rather, it’s simply because there is a way of thinking *about* the experience of seeing red that is unavailable to her while still in the room.

Before Mary comes out of the room, she lacks a phenomenal concept of the experience of seeing red. She could always think about the experience using her old material concepts, but not with any phenomenal concept. This is why she did not know that *seeing red* = *THAT experience* (where this is to be understood as using a material concept on the left-hand side and a phenomenal concept on the right).

The crucial feature of phenomenal concepts, for the purposes of this argument, is that they are *experience-dependent*: the concept’s acquisition depends on its
end p.126

possessor having previously undergone the experience it refers to. This is why she doesn’t “know what seeing red is like” before she comes out of the room. She needs to see red in order to acquire the conceptual wherewithal to think *seeing red* = *THAT experience*.

The reason that Mary’s new concept depends on experience is that it requires a sensory template, and her acquisition of this template depends on her visual system’s having been activated previously by some red surface. Of course, this is a contingent feature of human beings. We can imagine beings who are born with the sensory templates that we acquire from color experiences (cf. Papineau 2002, sec. 2.8). But humans are born with few, if any, sensory templates; rather, they must acquire them from previous experiences. (If humans were born with the sensory templates activated by red surfaces, then physicalists could not answer the knowledge argument by saying that Mary needs a red experience to acquire a phenomenal concept of red. But if humans were like that, then physicalists need not answer the knowledge argument in the first place: since Mary would have a phenomenal concept of red before she left her room, she would already be in a position to know that *seeing red* = *THAT experience*.)

I Am Not Now Having or Imagining That Experience

It is worth distinguishing the experience-dependence of phenomenal concepts from the use-mention feature discussed above. Even though normal examples of phenomenal concepts, such as the one Mary acquires on leaving her room, have both the experience-dependent and use-mention features, there is space in principle for concepts that are phenomenal in the experience-dependent sense but that don't use experiences to mention themselves. Indeed, I would argue that this is not just an abstract possibility—there are actual concepts that display experience dependence but not the use-mention feature. To see why, recall the earlier discussion of perceptually derived concepts. These derived concepts involved the creation of some nonsensory file to house the information associated with some perceptual concept, and they made it possible to think about perceptible entities even when those entities were not being perceived or perceptually imagined. Analogously, we can posit a species of “phenomenally derived concept.” Suppose someone starts off, like Mary, by thinking phenomenally, using a sensory template instilled by previous experiences. But then she creates a nonsensory file to house the information that has become attached to that template, which will henceforth allow her to think about the experience without any sensory activation. I say she now has a phenomenally *derived* concept. Exercises of this concept won't activate the experience it mentions, and so this concept will fail to satisfy the use-mention requirement. But this phenomenally derived concept will still satisfy the experience-dependence requirement because its creation depends on a prior phenomenal concept that in turn depends on previous experiences.

The possibility of phenomenally derived concepts offers an answer to an objection raised at the beginning of this chapter. There I said that standard accounts of phenomenal concepts seem to imply that any exercise of a phenomenal concept demands the presence of the experience it refers to or an imaginatively re-created exemplar thereof. However, this seems too demanding. Surely someone like Mary can use her new concept to think truly, *I am not now having THAT experience (nor re-creating it in my imagination)*. Yet this should be impossible, if any exercise of her phenomenal concept does indeed require the relevant experience or its imaginative re-creation.

My response to this objection is that Mary thinks the problematic thought with the help of a phenomenally derived concept.¹³ She starts with a phenomenal concept based on some sensory template, but her subsequent creation of a nonsensory file allows her to think about the relevant experience without activating that template. She thinks, *I am not now having or imagining that experience*. And because she is using a phenomenally derived concept, what she thinks may well be true.

Some readers might wonder whether it is really appropriate to say that Mary is here exercising a *phenomenal* concept. After all, if this concept is realized nonsensorily, then why is it any more “phenomenal” than the general run of ordinary concepts? In particular, would we want to say that someone knows “what it is like” to see something red merely by virtue of her thinking *seeing red* = *THAT experience* where the right-hand side deploys a phenomenally derived concept?

Well, I have no principled objection if someone wants to withhold the description “phenomenal” on these grounds. But note that this move is not available to someone who wants to press the objection at hand, which after all is precisely that there seems room for a thinker to exercise a “phenomenal” concept while not having any version of the experience referred to. For this objection to make any sense, “phenomenal” cannot be

understood as requiring sensory realization per se. Rather, it has to be understood simply as standing for those concepts whose acquisition depends on undergoing the relevant experience. And in this sense of “phenomenal,” phenomenally derived concepts do explain how someone can think phenomenally without having any version of the corresponding experience.

Semantic Stability and A Posteriori Necessity

Let me now turn to antimaterialist arguments that depend on modal considerations. The best known of these is Kripke's argument against the identity theory in *Naming and Necessity* (1980). But before addressing this, I would like to consider a different modal argument, which I shall call “the argument from semantic stability.” As it turns out, both these arguments can be blocked by appealing to the use-mention feature of phenomenal concepts. But the way this works is rather different in the two cases.
end p.128

The argument from semantic stability hinges on the fact that type-B physicalists take identity claims such as *nociceptive-specific neuronal activity* = *pain* (where the right-hand side uses a phenomenal concept) to be a posteriori necessities. The distinctness of the concepts on either side of the identity claim means that there is no question of knowing such claims a priori. Even after she acquires both concepts, Mary still needs empirical information to find out that *pain* is the same experience as *nociceptive-specific neuronal activity*. In this respect, type-B physicalists take mind-brain identity claims to be akin to such familiar a posteriori necessities as *water* = H_2O , *lightning* = *electric discharge*, or *Hesperus* = *Phosphorus*.

The objection to type-B physicalism is that phenomenal mind-brain identities cannot possibly be akin to these familiar a posteriori necessities, because a posteriori necessity is characteristically ascribable to “semantic instability,” but phenomenal concepts are semantically stable.¹⁴

Let me unpack this. Note first that, in all the examples of familiar a posteriori necessities listed above, the referential value of at least one of the concepts involved—*water*, *lightning*, *Hesperus* (and indeed *Phosphorus*)—depends, so to speak, on how things actually are. If it had turned out that XYZ and not H_2O is the colorless liquid in rivers, then *water* would have referred to XYZ. If it had turned out that some heavenly body other than Venus is seen in the early morning sky, then *Hesperus* would have referred to that other heavenly body. And so on.

This observation suggests the hypothesis that a priority and necessity only come apart in the presence of semantically unstable concepts. On this hypothesis, claims formulated using semantically *stable* concepts will be necessary if and only if they are a priori. Certainly there are plenty of concepts that seem to be stable in the relevant sense, that is, whose referents seem not to depend on the actual facts. Physical concepts such as *electron* or H_2O seem to be like this, as do such everyday concepts as *garden* or *baseball*. And if we stick to claims involving only such stable concepts (*electrons are*

negatively charged, say, or *baseball is a game*), then it does seem plausible that these claims will be necessary if and only if they are a priori.

The general idea here is that necessities will be a posteriori only when you are ignorant of the essential nature of some entity you are thinking about. If your concepts are transparent to you, if their real essence coincides with their nominal essence, so to speak, then you will be able to tell a priori whether claims involving them are necessary. But with semantically unstable concepts, we need empirical information to know what they refer to and so to ascertain whether a necessary proposition is expressed. To take just the first example above, it takes empirical work to discover that H_2O is the referent of *water* and so that *water* = H_2O is necessarily true.

The claim is thus that a posteriori necessity always turns on the presence of concepts whose reference depends on the actual facts; correlatively, if we keep away from such concepts, then necessity and a priority will always go hand in hand.
end p.129

If this claim is accepted, it is hard to see how phenomenal mind-brain identity claims such as *pain* = *nociceptive-specific neuronal activity* could be true. For the phenomenal concepts like *pain* do *not* seem to be semantically unstable. There seems little sense to the idea that it could have turned out, given different empirical discoveries, that *pain* referred to something other than its actual referent. But then, given the general thesis that a posteriori necessity requires semantic instability, it follows that *nociceptive-specific neuronal activity* = *pain* cannot be an a posteriori necessity. Since we don't need any empirical information to know what *pain* refers to, we must already know what proposition the claim *nociceptive-specific neuronal activity* = *pain* expresses, solely by virtue of our grasp of the concept *pain*, and so ought to be able to tell a priori that this claim is true if it is. But we can't tell a priori that it's true. So it can't be true.

In the face of this argument, type-B physicalists need to deny that a posteriori necessities require semantically unstable concepts. There seems no doubt that phenomenal concepts are semantically stable. And it is constitutive of type-B physicalism that phenomenal mind-brain identities are a posteriori. So the only option left is to insist that these identities are a posteriori necessities that involve no semantic instability.

Opponents will ask whether phenomenal mind-brain identities are the only such cases. If they are, then the type-B physicalist can be charged with making an unacceptably ad hoc move. Type-B physicalists would seem to be guilty of special pleading if the connection between a posteriori necessity and semantic instability holds good across the board except in cases that involve phenomenal concepts.

One obvious way for type-B physicalists to respond to this charge is to seek other examples of a posteriori necessities that do not involve semantic instability. Obvious possibilities are identities involving proper-name concepts that (unlike *Hesperus* or *Phosphorous*) do not have their references fixed by salient descriptions (*Cicero* = *Tully* say), or identities involving perceptual concepts (such as *reflectance profile* Φ = *red* where the right-hand side uses a perceptual color concept).

However, opponents of type-B physicalism will deny that these a posteriori necessities are free of semantic instability. Maybe the concepts involved don't have their references fixed by salient descriptions. But the opponents will insist that this is not the only way for

concepts to be semantically unstable and that more careful analyses of semantic stability will show that proper-name and perceptual concepts are indeed unstable, whereas phenomenal concepts are not, and are thus still anomalous among concepts that enter into a posteriori necessities.¹⁵

I remain to be persuaded by this charge of anomalousness. However, I shall not dig my heels in at this point. Rather, let me concede, for the sake of the argument,
end p.130

that phenomenal mind-brain identities are indeed anomalous because they don't involve any semantically unstable concepts. I don't accept that this means that these identities cannot be true. Rather, I say that phenomenal mind-brain identities are anomalous because phenomenal concepts are very peculiar. More specifically, phenomenal concepts have the very peculiar feature of *using* the experiences they refer to. When we reflect on this, we will see that it is unsurprising that identities involving phenomenal concepts should be unusual in combining semantic stability with a posteriori necessity.

The underlying antiphysicalist thought, recall, was that semantic stability goes hand in hand with knowledge of real essences; conversely, if thinkers are ignorant of real essences, they must be using unstable concepts. The complaint about type-B physicalism, then, is that it requires the possessors of phenomenal concepts like *pain* to be ignorant of the real physical essence of pain, even though the concept *pain* is manifestly stable. The antiphysicalists thence conclude that *pain* must refer to something nonphysical, something with which the possessors of the concept are indeed directly acquainted.

But type-B physicalists can respond that, however it is with other concepts, this combination of semantic stability and ignorance of essence is just what we should expect, given the use-mention feature characteristic of phenomenal concepts. Even if phenomenal concepts don't involve direct knowledge of real essences, they will still come out semantically stable, for the simple reason that the use-mention feature leads us to think of the referent as “built into” the concept itself. Since the concept uses the phenomenal property it mentions, this alone seems to eliminate any conceptual or metaphysical space wherein *that* concept might have referred to something different.

I said above that I remained to be persuaded that phenomenal concepts are distinguished from proper-name and perceptual concepts in uniquely displaying this combination of semantic stability and ignorance of essence. Still, at an intuitive level we can see why phenomenal concepts should appear special in this way. When we think of a proper-name concept, such as *Cicero*, or a perceptual concept, such as *red*, we seem able to make intuitive sense of scenarios where the reference-fixing facts are different, that is, where the concept *Cicero* names some other person than Cicero, or where the perceptual observational concept *red* refers to some different surface property. But we don't seem able to do this with phenomenal concepts—there don't seem to be any scenarios whose actuality would make *pain* refer to something other than pain. This is because we think of phenomenal concepts as essentially *using* the very phenomenal properties that are being referred to.

end p.131

This seems to leave no room for the idea that a given phenomenal concept could have referred to some property other than the one it does refer to, if the facts had turned out differently. As long as it remains the same phenomenal concept, its exercises will involve the same phenomenal property—and then, since it mentions whichever phenomenal property it uses, it will refer to that property, however the actual facts turn out.¹⁶ In a sense, phenomenal concepts are too close to their referents for it to seem possible that those same concepts could refer to something else. With other concepts that enter into a posteriori identities, including proper-name and perceptual concepts, we can imagine the “outside world” turning out in such a way that they referred to something other than their actual referents. Some other person might have turned out to be the historical source of my *Cicero* concept, some other physical property might have turned out to answer to my *red* concept, and so on. But in the case of phenomenal concepts, the referent seems to be part of the concept itself, leaving no room for any such possibility.¹⁷ If this is right, then the semantic stability of phenomenal concepts provides no reason to think that they must refer to nonphysical properties with which their possessors are directly acquainted. For the use-mention feature of phenomenal concepts yields an independent explanation of why they should be semantically stable, even while their possessors remain ignorant of the real physical essences of their referents.

Kripke's Original Argument

Let me now turn to Saul Kripke's original argument against the mind-brain identity theory. There are significant differences between this and the argument from semantic stability. Kripke doesn't seem to be committed to the thesis that necessity comes apart from a priority only in the presence of semantic instability. Kripke's paradigm cases of a posteriori necessities involve names whose reference is determined in line with his causal theory of reference (*Cicero* = *Tully*), and there is nothing in Kripke to suggest that this a posteriority demands that the referential values of these names must depend on the actual facts. From a Kripkean point of view, there is nothing special here that needs explaining—a posteriori necessities are simply a natural consequence of the nondescriptive way reference is determined for normal names.

Kripke's argument hinges not on the a posteriority of the physicalist's identity claims, but on their *apparent contingency*. Kripke has no complaint about the a posteriority of such claims as *pain* = *nociceptive-specific neuronal activity* a posteriori necessities are par for the course, from his point of view. What Kripke takes to be problematic about these mind-brain identities, rather, is that it seems that they *might have been* false: intuitively we feel that there are possible worlds—zombie worlds, say—where nociceptive-specific neuronal activity is not identical to pain. Now there is nothing per se incoherent in the idea of a necessary identity that appears contingent. For example, we can make sense of the idea that *Hesperus* = *Venus* might have been false by construing this identity claim as saying that the heavenly body that appears in the morning is Venus—something which might well have been otherwise. But now—and this is Kripke's point—this way of

explaining the appearance of contingency *does* require you to construe the relevant referring term as semantically unstable, for it demands that you read the term in a way that makes it come out referring to something different if the actual facts are different. Yet *pain* cannot be read in such a way, which means, says Kripke, that the physicalist has no satisfactory way of explaining the apparent contingency of phenomenal mind-brain identities.

So Kripke gets to the same place as the argument from semantic stability, but he gets there from a different starting point. However, the differences between the two arguments mean that Kripke's argument demands a rather different response from the physicalist. It is no good to reply to Kripke that a posteriori necessity is consistent with semantic stability, for he agrees about this, and indeed will allow that there are nonphenomenal examples of this combination (*Cicero = Tully*). What he insists on, however, and this is a different point, is that apparently contingent truths are inconsistent with semantic stability, and that physicalists therefore have no way of explaining the apparent contingency of the mind-brain identities they endorse.

In response to the argument for semantic stability, I denied that a posteriori necessity required semantic instability, at least in the case of mind-brain identities. However, I don't think that there is any corresponding room to deny that apparently contingent truths must involve semantic instability. If a claim can be understood as actually true yet possibly false, then some of the concepts involved must shift reference. Given that the concepts in phenomenal mind-brain identity claims are all semantically stable, this leaves physicalists with one option: deny that their phenomenal mind-brain identities are apparently contingent.

This may seem all wrong. Isn't it agreed on all sides that we can cogently conceive of zombies (even if they aren't really possible), and therewith that mind-brain identities at least *seem* possibly false? So what room is there for the physicalist to deny that mind-brain identities are apparently contingent?

Although I agree that physicalists are compelled to allow that phenomenal mind-brain identities seem possibly false, this isn't to say that they seem *contingently true*.

Contingent truth requires not only falsity in some possible worlds, but also truth in the actual world. And it is specifically this combination that generates the need for semantic instability, to give a nonactual referent that falsifies the claim in some possible world. But there is another way for an identity claim to seem

end p.133

possibly false, namely, for it simply to seem *false*. And in that case there is nothing to require semantic instability.

Let me go more slowly. Consider people who think that Cicero is actually different from Tully. To them the claim *Cicero = Tully* will of course seem possibly false, because it will seem necessarily false. But nothing here demands that we understand them as thinking of either of these names in a semantically unstable way, as possibly referring to something other than their actual referents. Such a construal is only called for when we have the combination of both apparent possible falsity and actual truth. If somebody thought that *Cicero = Tully* is true, but *might have been false*, then we must construe them as thinking *the greatest Roman orator = Tully*, or some such, to explain how they

have room, so to speak, for the thought that a necessary truth might have turned out to be false. But there's no need to read them this way if they simply think that Cicero and Tully are *actually* different.

So my suggestion is that physicalists should say that mind-brain identities strike us just as *Cicero = Tully* strikes people who think Cicero and Tully are different people. They seem non-necessary simply because they seem false. Zombies seem possible simply because pain seems actually distinct from nociceptive-specific neuronal activity. From this point of view, Kripke has misdescribed the crucial zombie intuition from the start. It's not an intuition of apparently contingent truth (some confused intuition that pain *could* come apart from nociceptive-specific neuronal activity in some other possible world) but simply a direct intuition of falsity (pain *is* different from nociceptive-specific neuronal activity in the actual world).^{18, 19}

The Intuition of Distinctness

Some readers may be wondering why this last point doesn't concede the antiphysicalist case to Kripke. My suggestion is that physicalists should explain our
end p.134

intuitions about the possibility of zombies by allowing that mind-brain identity claims strike us as false. But isn't this tantamount to denying physicalism?

Not necessarily. I say physicalists should allow that physicalism *seems* false, not that it *is* false. That is, physicalists should maintain that we have an *intuition* of mind-brain distinctness but that this intuition is mistaken.

This is by no means ad hoc. It seems undeniable that most people have a strong intuition of mind-brain distinctness—an intuition that pains are something extra to brain states, say. This intuition is prior to any philosophical analyses of the mind-brain relation, and indeed persists even among those (like me) who are persuaded by those analyses that dualism must be false. Given this, it is a requirement on any satisfactory physicalist position that it offer some explanation of why we should all have such a persistent intuition of mind-brain distinctness even though it is false. Physicalists need to recognize and accommodate the intuition of distinctness, quite apart from the fact that they need it to deal with Kripke's argument.

There are a number of possible ways of explaining away the intuition of distinctness, especially for physicalists who recognize phenomenal concepts. I favor an explanation that hinges on the use-mention feature of phenomenal concepts, an explanation I have called “the antipathetic fallacy.”²⁰

Suppose you entertain a standard phenomenal mind-brain identity claim such as *pain = nociceptive-specific neuronal activity*, deploying a phenomenal concept on the left-hand side and a material concept on the right. Given that the phenomenal concept *uses* the experience it mentions, your exercise of this concept will depend on your actually having a pain, or an imagined re-creation thereof. Because of this, exercising a phenomenal concept will *feel* like having the experience itself. The activity of thinking phenomenally *about* pain will introspectively strike you as involving a version of the experience itself.

Things are different with the exercise of the material concept on the right-hand side. There is no analogous phenomenology. Thinking of nociceptive-specific neuronal

activity doesn't require any pain-like feeling. So there is an intuitive sense in which the exercise of this material concept "leaves out" the experience at issue. It "leaves out" the pain in the sense that it doesn't activate any version of it.

It is all too easy to slide from this to the conclusion that, in exercising such a material concept, we are not thinking *about* the experiences themselves. After all, doesn't this material mode of thought "leave out" the experiences in a way that the phenomenal concept does not? And doesn't this show that the material concept simply doesn't refer to the experience denoted by our phenomenal concept of pain?

This line of thought is terribly natural. (Consider the standard rhetorical ploy: "How could *pain* arise from mere neuronal activity?") But of course it is a fallacy. There is a sense in which material concepts do "leave out" the feelings. Uses of them do not in any way activate the experiences in question, by contrast with uses of phenomenal concepts. But it simply does not follow that these material concepts
end p.135

"leave out" the feelings in the sense of failing to refer to them. They can still refer to the feelings, even though they don't activate them.

After all, most concepts don't use or involve the things they refer to. When I think of being rich, say, or having measles, this doesn't in any sense make me rich or give me measles. In *using* the states they refer to, phenomenal concepts are very much the exception. So we shouldn't conclude on this account that material concepts, which work in the normal way of most concepts, in not using the states they refer to, fail to refer to those states.

Fallacious as it is, this line of thought still seems to me to offer a natural account of the intuitive resistance to physicalism about conscious experiences. This resistance arises because we have a special way of thinking about our conscious experiences, namely, by using phenomenal concepts. We can think about our conscious experience using concepts to which they bear a phenomenal resemblance. And this then creates the fallacious impression that other nonphenomenal ways of thinking about those experiences fail to refer to the felt experiences themselves.²¹

Chalmers on Type-B Physicalism

Chalmers's Dilemma

David Chalmers has recently mounted an attack on the whole type-B physicalist strategy of invoking phenomenal concepts to explain the mind-brain relation (see chap. 9, this volume). He aims to present type-B physicalists with a dilemma. Let C be the thesis that humans possess phenomenal concepts. As Chalmers sees it, type-B physicalists require both (a) that C explains our epistemic situation with respect to consciousness and (b) that C is explicable in physical terms. However, Chalmers argues that there is no version of C that satisfies both these desiderata—either C can be understood in a way that makes it

physically explicable or in a way that allows us to explain our epistemic situation, but not both.

To develop the horns of this dilemma, Chalmers asks the physicalist whether or not C is conceptually guaranteed by the complete physical truth about the universe, P. That is, is P and not-C conceivable?

Suppose the physicalist says that this combination *is* conceivable. This makes the existence of phenomenal concepts conceptually independent of all physical claims. But then, argues Chalmers, all the original puzzles about the relation between the brain and phenomenal states will simply reappear as puzzles about the relation between the brain and phenomenal concepts themselves. So Chalmers holds that on this option, C fails to be physically explicable.

The other option is for the physicalist to say that P and not-C is *not* conceivable. On this horn, claims about phenomenal concepts will not be conceptually independent of P (for example, suppose that phenomenal concepts are conceived physically
end p.136

or functionally). This would mean that zombies (conceivable beings who are physically identical to us but lack consciousness) would be conceived as having phenomenal concepts. However, Chalmers maintains, we don't conceive of zombies as epistemically related to consciousness as we are (after all, they are conceived as not having any consciousness). This argues that something more than C is needed to explain our peculiar relation to consciousness. So Chalmers holds that on this option, C fails to explain our epistemic situation.

So either P and not-C is conceivable, or it isn't. And on neither option, argues Chalmers, is C both physically explicable and explanatory of our epistemic situation.

The Dilemma Embraced

Far from viewing Chalmers as offering a nasty choice, I am happy to embrace both horns of his dilemma. I say that we can conceive of phenomenal concepts in a way that makes them conceptually independent of the physical facts *and* conceive of them in a way that doesn't make them conceptually independent. Moreover, I think that *both* these ways of thinking about phenomenal concepts allows phenomenal concepts to be simultaneously physically explicable and explanatory of our epistemic situation.

It is the use-mention feature of phenomenal concepts that allows them to be thought of in two different ways. Exercises of phenomenal concepts involve versions of the phenomenal states they refer to. Given this, thinking about phenomenal concepts requires us to think of the phenomenal states that they use. But according to type-B physicalism, these used phenomenal states, like phenomenal states in general, can be thought of in two different ways: phenomenally and nonphenomenally. So we can think about (first-order) phenomenal concepts phenomenally, using (second-order) phenomenal concepts to think about the phenomenal states involved, or we can think about them nonphenomenally, conceiving the involved phenomenal states in, say, physical or functional terms. Since the (second-order) phenomenal concepts used on the first option, like all phenomenal concepts, will be a priori distinct from any physical or functional concepts, the first way of thinking about (first-order) phenomenal concepts will make P and not-C conceivable,

whereas the second way of thinking about (first-order) phenomenal concepts, *as* physical or functional, will make P and not-C inconceivable.²²

However, for a type-B physicalist, these two ways of thinking still refer to the same entities: (first-order) phenomenal concepts. And these entities will have the same nature and cognitive role, however they are referred to. So the way they are referred to ought to make no difference to whether they are physically explicable and explanatory of our epistemic situation. They should satisfy these two desiderata in any case. Let me now show that they do.

The First Horn

On the first horn, we think about first-order phenomenal concepts by using second-order phenomenal concepts. That is, we note that exercises of first-order phenomenal concepts involve uses of phenomenal states, and when we think of the phenomenal states involved, we do so using second-order phenomenal concepts.

The problem on this horn, according to Chalmers, relates to the epistemic and explanatory gap between physical and phenomenal claims. Chalmers views this gap as making a strong case for dualism, a case that type-B physicalists seek to block by showing that the existence of this gap can be explained in terms of certain features of phenomenal concepts. However, even if invoking phenomenal concepts can succeed in explaining the original gap between physical and phenomenal claims, Chalmers argues that physicalists will now need to explain away a new gap between physical claims and claims about the possession of phenomenal concepts. For, after all, on this horn of the dilemma, they agree that P and not-C is conceivable, that is, that the physical facts do not conceptually necessitate claims about first-order phenomenal concepts.

The obvious physicalist response is to argue that they can explain this new gap in just the way that they explained the original one. If we are conceiving of first-order phenomenal concepts using second-order phenomenal concepts, then of course there will be a conceptual gap between physical claims and claims about phenomenal concepts. Still, if the original gap could be “explained in terms of certain distinctive features of [first-order] phenomenal concepts,” as Chalmers is allowing for the sake of the argument, why can't the new gap be explained in terms of the same features of second-order phenomenal concepts?

Chalmers objects that this explanation-repeating move will be either regressive or circular. But it is not obvious to me why this should be so. In particular, there doesn't seem to be anything regressive or circular in repeating the explanatory use that I have made of phenomenal concepts, as I shall show in a moment.

I suspect that Chalmers's charge of regression or circularity reflects the very high demands he is placing on type-B explanations of the conceptual gap. For the most part he leaves it open how such explanations might go, being happy to conduct his argument on an abstract level. But just before his charge of regression or circularity, he does propose one possible explanation of the original gap, suggesting that type-B physicalists might say that phenomenal concepts give their possessors a distinctive kind of direct acquaintance with their referents, of a kind that “one would not predict from just the

physical/functional structure of the brain.” I agree that *this* kind of explanation is going to get type-B physicalists into trouble, but not because it becomes regressive or circular when it is repeated at the higher level. Rather, the trouble comes because it is unacceptable at the outset for a physicalist to posit distinctive semantic powers of direct reference that correspond to nothing identifiable in physical or functional terms.
end p.138

Still, this doesn't mean that *any* type-B physicalist explanation of the conceptual gap is going to run into trouble. There is no question here of cataloging all the different ways in which different type-B physicalists have appealed to phenomenal concepts in order to account for various “gaps.” Let me simply remind readers of some of the things I said earlier and show that there is nothing regressive or circular about saying the same things about the relation between physical claims and phenomenally conceived claims about phenomenal concepts.

Like all type-B physicalists, I take the existence of first-order phenomenal concepts to imply that first-order phenomenal mind-brain identity claims are a posteriori. In response to the challenge that these claims are unique among a posteriori necessities in not hinging on semantically unstable concepts, I argued that such uniqueness is adequately explained by the use-mention feature of phenomenal concepts. This feature explains why we take it that phenomenal concepts will refer to the same referent whatever the facts, even though phenomenal concepts are arguably unlike other semantically stable concepts in not requiring transparent knowledge of the essential nature of their referents. As to the feeling that, even after this has been said, there remains something disturbingly unexplanatory about phenomenal mind-brain identities, my view is that this feeling doesn't stem from any semantic or epistemic peculiarity of these identities, but simply from the prior “intuition of distinctness” that militates against our believing these identities to start with. (To the extent we embrace this intuition, then of course we will feel a real “explanatory gap,” for we will then want some causal explanation of why the physical brain should “give rise to” the supposedly separate phenomenal mind.)²³ As to the source of the intuition of distinctness, I explained this by once more appealing to the use-mention feature of phenomenal concepts, and the way it makes us think that nonphenomenal modes of thought “leave out” the phenomenal feelings.

Now I don't see why I can't simply say all these things again, if Chalmers challenges me to account for the extra gap between P and C that arises when we are thinking of first-order phenomenal concepts in second-order phenomenal terms. Second-order phenomenal claims identifying the possession of first-order phenomenal concepts with physical states will be a posteriori necessities. If these are held to be unusual among a posteriori necessities in not involving semantic instability, I can point out that the use-mention feature of second-order phenomenal concepts explains why this should be so. If it is felt that, even after this has been said, there seems to be something disturbingly unexplanatory about claims identifying the possession of first-order phenomenal concepts with physical states, I attribute this to a higher level “intuition of distinctness” that arises because physical/functional ways of thinking about first-order phenomenal concepts seems to “leave out” the feelings which are present when we think about first-order phenomenal concepts using second-order phenomenal concepts.

end p.139

In short, just as the use-mention feature of first-order phenomenal concepts accounts for any peculiarities of the conceptual gap between physical/functional claims and first-order phenomenal claims about phenomenal properties, so does the use-mention feature of second-order phenomenal concepts account for any similar peculiarities in the gap between physical/functional claims P and second-order phenomenal claims C about the possession of phenomenal concepts.

Of course, Chalmers may now wish to ask about the relationship between physical/functional claims P and third-order phenomenal claims C about the possession of second-order phenomenal concepts. But I am happy to go on as long as he is.

The Second Horn

Let me now turn to the other horn of Chalmers's dilemma. Here P and not-C is *not* conceivable. We conceive of phenomenal concepts in physical/functional terms, and so we conceive of zombies as sharing our phenomenal concepts by virtue of sharing our physical/functional properties.

Chalmers's worry on this horn is that phenomenal concepts so conceived will fail to explain our epistemic relationship to consciousness. For we don't conceive of zombies as epistemically related to consciousness as we are, even though (on this horn) we conceive them as sharing our phenomenal concepts. So something more than phenomenal concepts seems to be needed to explain our peculiar epistemic relation to consciousness.

To rebut this argument and show that a physical/functional conception of phenomenal concepts allows a perfectly adequate account of our epistemic relation to consciousness, I need to proceed in stages. To start with, observe that none of the points I have made about our relationship to consciousness demands anything more than a purely physical/functional conception of phenomenal concepts. To confirm this, we can check that all these points would apply equally to zombies, conceived of as having physical/functional phenomenal concepts but no inner phenomenology. Note that the zombies' "phenomenal" concepts (the scare quotes are to signal that we are not now conceiving of these concepts as involving any phenomenology) will be just as *experience-dependent* as our own. Zombie Mary will need to come out of her room to acquire a "phenomenal" concept of red experience, and when she does, she will acquire some new non-indexical knowledge: she will come to know that *seeing red* = *THAT experience* (where this is to be understood as using a material concept on the left-hand side and her new experience-dependent "phenomenal" concept on the right-hand side). Moreover, this kind of knowledge is arguably unusual insofar as it lays claim to an a posteriori necessity but doesn't display the semantic instability characteristic of such claims. Still, zombie type-B physicalists can invoke the *use-mention* feature of zombie Mary's new "phenomenal" concept to explain why that concept should come out as semantically stable even though its possessors can be ignorant of the essential nature of its referent. Not that the zombie type-B physicalists are likely to have things all their own

way, for they will also have to contend with the zombie “intuition of distinctness”: zombies who reflect on the nature of their “phenomenal” brain-mind claims might well note (using second-order “phenomenal” concepts) that the left-hand sides “leave out” a end p.140

mental property that is used on the right-hand sides, and conclude on this basis that non-“phenomenal” concepts don't mention the same mental properties that are mentioned by “phenomenal” concepts. Still, zombie type-B physicalists can point out that this is a confusion, engendered by the peculiar use-mention feature of zombie “phenomenal” concepts.

All in all, then, everything I have said about our own epistemic relation to our conscious states will be mirrored by the zombies' relation to their corresponding states. I take this symmetry with zombies to show that our own relationship to consciousness can be perfectly adequately explained using a physical/functional conception of phenomenal properties. But Chalmers urges that the comparison cuts the other way. Maybe, he allows, we can suppose that zombies have states to which they stand in the same sort of epistemic relation that we have to consciousness. But we mustn't forget, he insists, that we are also conceiving of zombies as beings who lack *our* inner life, who have no phenomenology. Given this, Chalmers argues, an explanation of mental life that works for zombies can't possibly explain our relation to our own conscious phenomenology.

At this point it will help to recall the basic type-B physicalist attitude to zombies. Since type-B physicalists hold that human consciousness is in fact physical, they contend that zombies are metaphysically impossible. Any being who shares our physical properties will therewith share our conscious properties; not even God could make a zombie. At the same time, type-B physicalists recognize that we have two ways of thinking about phenomenal properties. This is why zombies are conceivable even though impossible. We can apply one way of thinking about phenomenal matters, but withhold the other—that is, we can think of zombies as sharing our physical/functional properties but as lacking our phenomenal properties phenomenally conceived.

Given this, there is no obvious reason why type-B physicalists should be worried that a physical/functional explanation of our epistemic relationship to consciousness will apply equally to zombies. Physical/functional duplicates of us will necessarily be conscious, just like us. True, our ability to think in phenomenal terms makes it possible for us also to conceive of these duplicates as lacking phenomenal properties, and thus as not being related to consciousness in the way that we are. But the fact that we can so conceive of zombies needn't worry the physicalist, who after all thinks that we are here conceiving an impossibility which misrepresents our actual relationship to consciousness. We can *imagine* beings who are physically/functionally just like us but who lack our inner life, but that doesn't mean that the physical/functional story is leaving something out, given that in reality our inner life isn't anything over and above the physical/functional facts. Let me conclude by turning to “silicon zombies.” Here things come out rather differently because type-B physicalism leaves open that silicon zombies are metaphysically as well as conceptually possible. Silicon zombies are possible beings who share our functional properties, if necessary down to a fine level of detail, but who are made of silicon-based materials rather than our carbon-based ones, and on that account lack our conscious

properties. As it happens, I am inclined to the view that conscious properties are identical with functional properties rather than strictly physical properties, and that silicon zombies are therefore metaphysically
end p.141

impossible, just like full-on zombies. However, nothing in type-B physicalism as I have presented it (nor indeed anything I have written about consciousness) requires this identification of conscious properties with functional rather than physical ones, so I am prepared to concede for the sake of argument that silicon zombies would lack conscious properties.

Now suppose further that the “physical/functional” conception of phenomenal concepts which defines the second horn of Chalmers's dilemma is in fact a functional conception. (This seems reasonable—all the nonphenomenal claims I have made about phenomenal concepts have hinged on their functional workings, not their physical nature.) Since silicon zombies are our functional duplicates, they will therefore have “phenomenal” concepts, functionally conceived, and these will mimic the operations of our own phenomenal concepts: silicon Mary will need to come out of her room to acquire a “phenomenal” concept of red “experience,” silicon subjects will suffer a “dualist intuition of distinctness,” and so on.

So my putative explanation of our own epistemic relationship to consciousness is mirrored by the matching relationship of the silicon zombies to their corresponding states, even though the silicon zombies are missing the crucial thing that we have: consciousness. And because we are dealing with silicon zombies rather than full-on physical duplicates, I can't say that this asymmetry is an illusion generated by our conceiving an impossibility, since by hypothesis the silicon zombies really wouldn't have the conscious properties that we humans possess. Unlike full-on duplicates, silicon zombies really do lack something that we have.

At this point, I think that type-B physicalists should bite the bullet and say that the thing that differentiates us from the silicon zombies doesn't make any difference to the explanatory significance of phenomenal concepts. We might be related to something different, but this doesn't mean that we enjoy some special *mode* of epistemological access to our conscious states that is not shared by the silicon zombies. After all, the silicon zombies' “phenomenal” concepts do successfully refer to a certain range of silicon mental properties—“schmonscious” properties—and type-B physicalists can say that the silicon zombies' “phenomenal” concepts relate them to these schmonscious properties in just the way that our own phenomenal concepts relate us to our conscious properties.

True, these schmonscious properties are not conscious ones, since by hypothesis we are supposing that consciousness requires carbon-based physical makeup. But this does not mean that there is any substantial explanatory asymmetry between the way our phenomenal concepts relate us to our conscious states and the way the zombies' “phenomenal” concepts relate them to their schmonscious states. (After all, note that silicon zombie philosophers can point out that *we* lack something that *they* have, given that we lack the silicon-based makeup required for schmonsciousness.)

Of course, if you are a dualist, like David Chalmers, or indeed like anybody who is still in the grip of the intuition of distinctness, then you will hold that there is some very special

extra property generated by carbon-based brains, and nothing corresponding in the silicon zombies. And you will think our introspective awareness relates us to this special extra property, and so must involve some special capacity that silicon zombies do not share. But physicalists reject any such special properties additional to physical/functional ones, and so they have no reason to
end p.142

think that our relation to our conscious properties is any different in kind from the silicon zombies' relation to their schmonscious properties. Phenomenal concepts, functionally conceived, provide a perfectly good explanation of both relationships.

Acknowledgments

Earlier versions of this chapter were read at the Tilburg workshop titled “Mind and Rationality” in August 2003, at the Jowett Society in Oxford in October 2003, and at the King's College Departmental Seminar in January 2004. I would like to thank all those who commented on those occasions.

References

- Balog, K. (2002). The “Quotational Account” of Phenomenal Concepts. Unpublished.
- Bealer, G. (2002). Modal Epistemology and the Rationalist Renaissance. In *Conceivability and Possibility*, ed. J. Hawthorne and T. Gendler: 71–125. Oxford: Oxford University Press.
- Chalmers, D. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
- Chalmers, D. J. (2002). Does Conceivability Entail Possibility? In *Conceivability and Possibility*, ed. J. Hawthorne and T. Gendler: 145–200. Oxford: Oxford University Press.
- Chalmers, D. J. (2003a). Consciousness and Its Place in Nature. In *The Blackwell Guide to the Philosophy of Mind*, ed. P. Stich and T. Warfield. Oxford: Blackwell. Reprinted in *The Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 247–72. New York: Oxford University Press, 2002.
- Chalmers, D. J. (2003b). The Content and Epistemology of Phenomenal Belief. In *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic: 220–72. Oxford: Oxford University Press.
- Chalmers, D. J., and Jackson, F. (2001) Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110: 315–61.
- Crane, T. (forthcoming). Papineau on Phenomenal Concepts. *Philosophy and Phenomenological Research*.
- Harman, G. (1990). The Intrinsic Quality of Experience. *Philosophical Perspectives* 4: 31–52. [Link](#)
- Horgan, T. (1984). Jackson on Physical Information. *Philosophical Quarterly* 34: 147–83. [Link](#)

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. Cambridge: MIT Press.

Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy* 83: 291–95.

 [Link](#)

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford: Oxford University Press.

Kripke, S. (1972). Naming and Necessity. In *The Semantics of Natural Language*, ed. G. Harman and D. Davidson. Dordrecht: Reidel. Reprinted as *Naming and Necessity*. Cambridge: Harvard University Press, 1980.

Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview. Revised version in *The Nature of Consciousness*, ed. by N. Block, O. Flanagan, and G. Güzelde: 597–616. Cambridge: MIT Press, 1997.

Melnyk, A. (2003). Contribution to symposium *Thinking about Consciousness*. Available at http://www.swif.uniba.it/lei/mind/forums/004_0003.htm
end p.143

Millikan, R. (1990). The Myth of the Essential Indexical. *Noûs* 24: 723–34.  [Link](#)

Millikan, R. (2000). *On Clear and Confused Ideas*. Cambridge: Cambridge University Press.

Papineau, D. (1993a). *Philosophical Naturalism*. Oxford: Blackwell.

Papineau, D. (1993b). Physicalism, Consciousness, and the Antipathetic Fallacy.

Australasian Journal of Philosophy 71: 169–83.  [Link](#)

Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press.

 [Link](#)  [OSO X-Reference](#)

Perry, J. (1979). The Problem of the Essential Indexical. *Noûs* 13: 3–12.  [Link](#)

Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press.

Prinz, J. (2000). A Neurofunctional Theory of Visual Consciousness. *Consciousness and Cognition* 9: 243–59.  [Link](#)

Prinz, J. (2002). *Furnishing the Mind*. Cambridge: MIT Press.

Stoljar, D. (forthcoming). The Argument from Diaphanousness. *New Essays in Philosophy of Language and Mind*, Canadian Journal of Philosophy, Suppl. Vol., ed. by M. Escudria, R. J. Stainton, and C. Viger.

Tye, M. (2003). A Theory of Phenomenal Concepts. In *Minds and Persons*, ed. A. O'Hear: 91–105. Cambridge: Cambridge University Press.

end p.144

eight Phenomenal Concepts and the Materialist Constraint

Joseph Levine

We begin with the following contrast. When considering the theoretical reduction of water to H₂O, we find the connection explanatory. We see why water is liquid at room temperature, transparent, turns to vapor under suitable circumstances, and the like, by

appeal to its chemical structure and other facts expressible in the vocabulary of physics and chemistry. The connection posited between water and H_2O in no way appears arbitrary to us.

On the other hand, when we consider the theoretical reduction of a phenomenal state, such as a visual experience, something important seems to be left unexplained. Though appeal to the neurological structure of the state (together with an account of the overall physical structure of the relevant portions of the nervous system) explains a lot about how various stimuli cause the visual experience and how the visual experience interacts with other states to cause both behavior and other cognitive states, the qualitative character of the experience—what it's like to have the experience—does not seem to be explained. The connection between the neurological description and our first-person conception of what it's like seems totally arbitrary. One feels that this neurological configuration could just as easily have gone with a bluish visual experience as a reddish one. In fact, for all we can tell, it could just as easily have gone with a state that was like nothing at all for the subject.

It is this contrast—this sense of arbitrariness that attends the psychophysical reduction as opposed to the sense of intelligibility that attends other theoretical reductions—that is the core problem that goes by the name of “the explanatory gap.” Though there are many different ways to illustrate this contrast—for instance, by appeal to the conceivability argument, Frank Jackson's (1982) case of Mary, and the open question argument—it's important not to lose sight of the core idea itself, namely the contrast. It seems to me that often in the literature this core idea has been lost amid the finer details concerning questions of modality, derivability, and identity.

To illustrate what I mean, consider the issue of derivability. The question of derivability arises once we link the explanatory gap to the conceivability argument. The conceivability argument proceeds from the premise that the existence of
end p.145

zombies cannot be ruled out a priori to the conclusion that they are metaphysically possible. The explanatory gap is implicated in the conceivability premise itself because, it is alleged, if we had an explanation of the phenomenal in terms of the physical, zombies wouldn't in fact be conceivable. Conceivability itself, as is clear from the statement of the argument above, has been interpreted as the absence of an a priori derivation of such statements as “This creature is having a reddish visual experience” from statements involving only physiological and functional vocabulary.

The issue of derivability has also come up in the discussion of Jackson's “knowledge argument.” The punch line of the argument is that when Mary finally leaves what Perry (2001) calls the “Jackson room” and is exposed to a red surface, she exclaims, “Oh, so *that's* what it's like to see red.” However, goes the argument, if phenomenal properties were really explicable in terms of their physiological correlates, she should have expected it to be like that. Based on all the theoretical knowledge she had accumulated in the Jackson room, she should have been able to predict what it would be like. This lack of ability to predict what it would be like is then interpreted to be a matter of her inability to come up with an a priori derivation of the relevant statement in “phenomenal” language from the theoretical descriptions couched in nonphenomenal language.

Once the question at issue has been transformed into one about a priori derivability, then the following dialectic takes hold. Defenders of materialism argue that in fact there is no relevant contrast between the standard cases of theoretical reduction and the psychophysical case. They argue that just as one cannot derive a priori statements of the form “This creature is having a reddish visual experience” from statements containing only nonphenomenal vocabulary, so too one cannot derive a priori statements such as “This cup is filled with water” from statements containing only vocabulary from chemistry and physics. If this absence of a priori derivability makes zombies conceivable, then so too is “zombie-H₂O” conceivable (i.e., H₂O that isn't water). Just as the conceivability of the latter does not throw any doubt on the metaphysical claim that water is H₂O—nor does it undermine the felt intelligibility and nonarbitrariness of the connection—so too the mere conceivability of zombies should throw no doubt on the claim that phenomenal properties are physical properties—or, more important, on the claim that the psychophysical link is fully intelligible.

Consider now the issue of identity. In presenting the core contrast underlying the claim that there is an explanatory gap between the phenomenal and the physical, I pointed to how we can explain many of the properties of water—such as its liquidity at room temperature—by appeal to its chemical structure (along with other relevant chemical facts, of course). I then noted that we seem unable to explain a phenomenal property, like the reddishness of a visual experience, by appeal to its underlying neurological structure. But of course there is an important difference here. In the water case, we look at some of water's properties and then seek an explanation in terms of the properties of H₂O. But in the psychophysical case, the *explanandum* at issue is the instantiation of the property that is alleged to be the very same property as that cited in the *explanans*: phenomenal reddishness, according to the materialist, *just is* the relevant neurological property. So of course there is no explanation, since identities are not really susceptible of explanation.

end p.146

Just ask not why water is liquid at room temperature, but why water is H₂O, to see how misleading the initial presentation of the core contrast was. After all, what is there to say in answer to why water is H₂O but that that's just the way it is? A similar answer, it is alleged, is appropriate in the psychophysical case.

Though these two materialist replies that focus on the issues of derivability and identity contain important insights, I think they still miss the crucial point, which is the core contrast with which we began. Suppose I agree that in fact one cannot derive a priori “water” talk from chemical talk precisely because the difference in vocabulary blocks the derivation. Suppose I also agree that identity claims, strictly speaking, do not require explanations. Still, the following contrast continues to stare us in the face. If after learning all the relevant chemistry—after learning how the molecular structure H₂O is responsible for all the superficial properties by which we identify water—someone were nevertheless to assert that she still didn't see how water could be H₂O, that the connection between being H₂O and being water seemed quite arbitrary, I believe we literally wouldn't understand what she was talking about. Yet, after learning all that Mary supposedly learns in the Jackson room, we understand quite well the feeling Mary has when she says, “Oh, so *that's* what it's like to see red!” There is a clear sense of

arbitrariness about it, a sense that it could easily have been some other way. If she were to follow up her exclamation with the question, “But why should it be like *that*?” we’d know what she means. The fact that we lack a priori derivability in both cases doesn’t remove the fundamental epistemic or cognitive contrast.¹

One may reply here that by insisting on this core contrast, I am refusing to acknowledge the crucial point just mentioned concerning the inexplicability of identities. A cognitive sense of “arbitrariness” is a symptom of having hit explanatory bedrock. When confronted with a question such as why the fundamental laws of physics are as they are—why, say, light travels the speed it does—it’s acceptable to answer that that’s just the way the world is. Yes, it is arbitrary, because there is no more fundamental phenomenon by appeal to which we can explain it. In effect, this is the property dualist answer to the explanatory gap. Why, the dualist says, does it seem arbitrary that this phenomenal property is correlated with that physical property? Her answer is that this reflects a basic law of nature. Appeal to the basic nature of a connection is a satisfactory explanation of its apparent arbitrariness.

Of course the materialist can’t appeal to a basic law connecting phenomenal and physical properties, for that would be to accept a form of dualism. But it appears she can still co-opt the “That’s just the way it is” response by appealing to the basic
end p.147

or brute nature of identity. When we reach identity claims, it is argued, we have also reached explanatory bedrock. Now I grant this, to an extent, but it doesn’t address the problem presented by the core contrast. After all, in both the water-H₂O and psychophysical cases we are dealing with identity claims, yet our cognitive assessment of the arbitrariness of the links is quite different. We need an account of this difference. Perhaps we can put the matter this way. We begin by identifying certain phenomenal states, such as visual experiences, with certain neurological states. We find that by doing so we can explain many of the properties of the phenomenal state, as mentioned above. We can explain, basically, its causal role. We then seek an explanation of its qualitative or phenomenal character. Why, we ask, is it like this to see red? Appeal to neurological properties doesn’t seem to answer the question. The neurological properties seem apt only for explaining causes and effects. But then the materialist makes the next move. You see, she says, the qualitative character *just is* one of the neurological properties. At this point, we intuitively balk. How can that be, we retort? They certainly don’t *seem* to be the same property. The idea seems unintelligible, in a way that the identity of water with H₂O doesn’t. So now we face, rather than the explanatory question, the non-identity question. Why, in this case, does it seem so bizarre to consider what is picked out by the one vocabulary to be the very same thing as what is picked out by the other vocabulary, when no such sense of bizarreness attends other theoretical reductions? Whether we think of it as an explanatory gap or a distinctness gap, the problem is really the same. Before turning finally to a discussion of phenomenal concepts, I want to pick up one loose thread from above. I noted earlier that the explanatory gap has been connected to the conceivability argument, to the knowledge argument, and also to the open question argument. I think a consideration of this last link will reinforce my argument concerning the relevance of the core contrast. By the “open question” argument, I have in mind the following. Suppose we are confronted with an alien species or an advanced robot. We

know everything there is to know about its internal workings. It turns out that its functional organization is, down to a fairly low level of implementation, very much like our own, though the physical mechanisms are different. Now we ask, is it conscious? And, if so, is what it is like for this creature to see red the same as what it is like for us? Contrast this case with the famous case of Twin Earth. On Twin Earth, we find a substance that behaves, at the macro level, exactly the way water behaves on Earth. We are happy to consider it water. However, after we conduct chemical tests on it, we see that its underlying structure, XYZ, is quite different from H₂O. Is it water nevertheless? Of course, those of us growing up in philosophy after the Kripke-Putnam revolution are prone to respond immediately that it isn't water. On the other hand, one can see how one might make the argument that it is water, that there are two kinds of water. Some might compromise and claim that it's really not determined in advance, that we just have to decide how we want to use the term "water" and whether or not we want to include XYZ in its extension.

Any of these three positions makes sense, I contend. We might have good reasons for picking one above the others as a better semantic hypothesis, but one can imagine arguments for all three. The point is, what all three positions have in
end p.148

common is the view that, after all the chemistry is done, what is left is a semantic issue. There is no real question of nonsemantic fact left open. All the relevant nonsemantic facts are revealed. This situation stands in stark contrast to how the psychophysical case described above at least appears. Whether or not the newly encountered creatures possess consciousness, and, if so, what it's like for them, does not seem at all a semantic question. What is left open, it seems quite clear, is a matter of genuine, nonsemantic fact, one that we haven't a clue how to go about determining. After all, what would it mean to "just decide" to call these creatures "conscious" or to call their phenomenal character when looking at red "reddish"?

I think the presence of an open nonsemantic question in the psychophysical case and its absence in the water case beautifully captures the core contrast. That there seems to be a genuine open question reflects the strong intuition that we are dealing with two properties here, the phenomenal one and the physical one, and that there is at best a brute, arbitrary connection between them. Though I don't think the falsity of the materialist identity claim is *entailed* by this intuition of distinctness, I do believe the materialist has a burden to explain its persistence. This is precisely what some materialists have attempted recently by appealing to the special properties of what are called "phenomenal concepts."

Most materialist attempts to explain the existence of an explanatory gap revolve around the idea of a phenomenal concept. Phenomenal concepts are concepts of phenomenal properties, the ones employed in standard first-person thoughts about one's conscious experience. So, for instance, when I wonder how the reddishness of my visual experience is explained by appeal to its physical properties, the concept of reddishness that is a constituent of that thought is a phenomenal concept. Though other concepts—those, say, that are expressed by vocabulary from neuroscience or computational psychology—might pick out the phenomenal property of reddishness, they don't qualify as phenomenal

concepts. Phenomenal concepts are quite special ways of conceiving of our sensory experiences, proprietary to the first-person point of view.

The general materialist strategy then is this. The initial puzzle concerns a certain cognitive state, our state of wondering how the physical properties of our sensory states explain their phenomenal properties. That we have such cognitive states is *prima facie* puzzling because the phenomenal properties in question just are the physical properties in question, so what's to explain? Obviously, they *seem* to be different properties, and the explanatory question makes sense to us. The answer to the puzzle presented by these cognitive states is to note that they involve, as constituents, two radically different kinds of concept: phenomenal concepts and, for want of a better term, nonphenomenal concepts. It is because we are conceiving of phenomenal properties via these two different kinds of concept that the explanatory question makes sense, that they seem so strongly to be about distinct properties. But the distinctness is all in the concepts, not the properties the concepts are concepts of.

Notice that the strategy just outlined does not merely appeal to the distinctness of the concepts involved, but to the alleged radical difference in kind between phenomenal and nonphenomenal concepts. This is crucial because we are able to see,
end p.149

with no special cognitive difficulty, how numerous properties and individuals that are picked out by distinct concepts might be the same thing. We started, after all, by noting such contrasts. So the critical move on behalf of the materialist is to provide an account of phenomenal concepts—and what's involved when we bring them together with nonphenomenal concepts within the same thoughts—that explains the unique cognitive features of this case. The attempts to provide just such an account are what I want to investigate in this chapter.

In the following sections, I will look at various ways of characterizing phenomenal concepts that are supposed to explain the presence of an explanatory gap. Though I base these characterizations loosely on the literature, none of them precisely corresponds to any one particular theorist, and some may not correspond to any. My purpose is to systematically survey the options for explaining the explanatory gap by appeal to the peculiar features of phenomenal concepts, not to take issue with any one such account.² Before beginning this survey of the options, I want to emphasize that any materialist proposal for explaining the gap must meet a condition I will call the Materialist Constraint: namely, that no appeal be made in the explanation to any mental property or relation that is basic. For instance, suppose a materialist argued that she could explain why there should be an explanatory gap (even though phenomenal properties were constituted by physical-functional properties) by appeal to some basic mental relation like acquaintance that held between subjects of experience and their brain states when conceiving of those states via phenomenal concepts. Acquaintance itself is not given a materialist explanation, but appealing to it, let us say, removes the mystery of the gap with respect to phenomenal properties. How appeal to acquaintance might do the job is not important for now (I will return to this later). The point is that it does the materialist no good to explain away the gap by violating her own doctrine—that is, by admitting into her ontology a mental relation that is basic. Thus, in our examination of the various

proposals to follow, it will be crucial to note that violation of the Materialist Constraint immediately disqualifies a proposal because it ceases to be a *materialist* explanation of the gap. Whatever it is that makes phenomenal concepts special, it must be possible to see how this feature can be implemented in a physical system by physical mechanisms. I begin with perhaps the simplest idea. Phenomenal concepts are representational primitives, and therefore thoughts containing them cannot be derived from thoughts that do not contain them. The mere difference in mental “vocabulary” between phenomenal language and nonphenomenal language explains why no purported explanation containing only nonphenomenal language in the explanans and phenomenal language in the explanandum can succeed.

This account is based on two principal ideas: that the explanatory gap is primarily a derivability gap, and that most ordinary lexicalized concepts possess

end p.150

analyses that in principle allow their elimination. As can be seen from the discussion above, both ideas are necessary to make this move work. I noted that one common reply to the conceivability argument is to maintain that metaphysical necessity can exist even in the absence of a priori derivability, and to point to the standard theoretical reductions as evidence of this. My counter-reply was to emphasize that we still need to explain the core contrast between the psychophysical case and the standard cases. Clearly, absence of derivability can't be the distinguishing factor, since we lack upward derivability in both cases. For these materialists, then, appeal to the fact that phenomenal concepts are representational primitives doesn't help, since it doesn't distinguish them from many nonphenomenal concepts.

However, I am now imagining a materialist who agrees that in general, most macro-level descriptions can be derived from the relevant micro-level descriptions. We do have upward derivability in the case of water, heat, and all the other standard cases of theoretical reduction. By admitting that upward derivability is normally present when there is upward metaphysical necessitation, the materialist can pin the core contrast between the psychophysical case and the standard cases on the presence or absence of upward derivability. Then, contrary to the antimaterialist's insistence that this lack of derivability reflects a metaphysical distinction in properties, the materialist maintains that it only reflects the primitive nature of phenomenal concepts. They *seem* to be about distinct properties, but that is only because of the distinctive character of the concepts by which we conceive them.

I present this option only to set it aside for now. I have two reasons. First, this position seems quite implausible on its face, and few materialists would endorse it. One has to buy the idea that only phenomenal concepts (along with logical, mathematical, and indexical concepts, and perhaps the concept of causation) are primitives, and that all of our other concepts are ultimately definable in terms of them. This sure sounds like the discredited doctrine of phenomenalism. Second, even if one did bite this bullet, I think there are independent reasons for thinking that mere lack of derivability is not the principal issue here. But to make that case we need to survey some of the other options first.

The first move was to account for the special character of phenomenal concepts by appealing to their status as conceptual primitives. This didn't work because many

nonphenomenal concepts are primitive as well. The next move also involves imputing a kind of primitiveness to phenomenal concepts, but this time it is not conceptual or representational but rather epistemic or judgmental. The idea is this. For most concepts, when we apply them in judgment—for example, judging that there's a dog in front of me, there's water in that glass, and so on—our application of the concept to an object depends on the application of other concepts. These other judgments serve as evidence for the judgment in question. So I judge that there's a dog in front of me because it appears to me that there's a dog in front of me; similarly for the judgment that there's water in that glass. Of course I needn't consciously go through this inference. But if someone were to challenge my claim about the dog or the water, my justification would certainly involve mentioning how things visually appear to me.

end p.151

Now, when it comes to phenomenal judgments—say, I'm having a reddish visual experience, I'm having a headache—there don't seem to be any epistemic liaisons of this sort to serve as evidence. I judge that I'm having a reddish visual experience because I am; the same for the headache. After all, what else could I point to? In these cases, the phenomenal states themselves seem directly to give rise to the judgments with no evidentiary intermediate.³ Notice that the phenomenal states do not themselves serve as evidence, since evidence already presupposes conceptualization. Rather, it's just that we are set up, when things work normally, to (be in a cognitive position to) judge that we are having a certain experience whenever we are.

Let's suppose that epistemic primitiveness really is a distinctive feature of phenomenal concepts. We still need to know why the fact that phenomenal concepts are applied in this epistemically primitive manner should give rise to the explanatory gap. How is that account supposed to go?

Perhaps the idea is this. The fact that most ordinary concepts possess epistemic liaisons of a type that phenomenal concepts lack helps to explain why accounts of underlying mechanisms in these cases yield satisfying explanations. For instance, when I identify something as water, I do this on the basis of its taste, visual appearance, feel, and the like. The chemical account of water not only provides a candidate for water's identity but also links that candidate to the mechanisms responsible for these other features by means of which I normally identify water. By thus embedding the concept of H_2O in a story that connects not just to the concept of water but also to all the other concepts of its epistemic liaisons, the sense that we have a genuinely explanatory account of water is generated. But if this is how the appeal to epistemic primitiveness is supposed to work, it doesn't in fact succeed. Though it may be true that phenomenal concepts can be applied in judgment without the application of other concepts in an evidentiary manner, this doesn't mean that phenomenal concepts are bereft of relevant links to other concepts. In particular, an account of the mechanisms underlying the production and interaction of phenomenal states would link up with many other concepts, especially nonphenomenal ones. We have a rich body of beliefs concerning the causes and effects of phenomenal states—composed of both phenomenal and nonphenomenal concepts in the very same cognitive states—and they constitute a ready set of explananda for theoretical accounts of phenomenal experience. Phenomenal concepts are well integrated with other concepts, as

the very facts mentioned in the paragraph above attest. After all, what are these other concepts that are connected to our application of the concept of water in judgment if not phenomenal concepts concerning the way water affects our sensory experience? So the mere fact that we are built to apply these representations without evidential intermediaries, and that they are activated only after the instantiation of the relevant phenomenal states, doesn't explain why thinking of phenomenal states by way of end p.152

them should make these states seem so arbitrarily connected to their neurological correlates.

I think there is a lesson to learn from the inadequacy of these first two accounts. What we're trying to explain is this. When we entertain these two different concepts of what is supposed to be the same property, we can't resist thinking of them as picking out distinct properties. We were trying to explain this phenomenon by appealing to a kind of cognitive isolation distinctive of one of the two concepts, the phenomenal one. But we saw that phenomenal concepts maintained links to nonphenomenal concepts, and though they may be primitive and unanalyzable, that didn't distinguish them from many nonphenomenal concepts. So it seems as if cognitive isolation isn't really the issue. Perhaps the answer lies in the relation between phenomenal concepts and their objects, phenomenal properties, rather than in the relation between phenomenal and nonphenomenal concepts. That is, maybe there's something special in the way phenomenal concepts represent phenomenal properties that accounts for this strong sense we have that when entertaining both a phenomenal and a nonphenomenal concept, we are dealing with distinct phenomena. Of course any model of that special relation between phenomenal concept and phenomenal property must respect the Materialist Constraint if it is going to defend materialism against the challenge of the explanatory gap. Further support for the idea that what's special about phenomenal concepts is the way they relate to their objects comes from another significant feature of phenomenal concepts that calls out for explanation: the fact that only subjects who have actually experienced the relevant phenomenal properties are capable of possessing the corresponding phenomenal concepts. For example, one of the main ideas underlying the Mary case is that one cannot employ a phenomenal concept unless one has personally instantiated the property it is a concept of. The reason Mary supposedly learns something new, is able only upon leaving the Jackson room to judge "So *that's* what it's like to see red," is that one can't employ, or even possess, phenomenal concepts until one has experienced the corresponding phenomenal states. Whereas one doesn't need to have instantiated constricting blood vessels in order to have a concept of them (the relevant nonphenomenal concept, that is), one does need to have experienced a headache to have that special first-person phenomenal concept of a headache. Let us refer to this feature of phenomenal concepts as the etiological constraint. Clearly there is something special in the way phenomenal concepts and their corresponding properties are related that explains the etiological constraint, and it stands to reason that whatever this special feature is, it also explains why there is an explanatory gap as well. In the next section we'll explore one way of trying to capture this special nature of the phenomenal concept-phenomenal property relation.

The model to be explored in this section incorporates the features of conceptual and epistemic primitiveness, but it adds a crucial element: phenomenal concepts are taken to involve an essential demonstrative component. I am going to use John Perry's (2001) discussion of Jackson's knowledge argument as my target here. He doesn't explicitly address the issue of the explanatory gap, but, as discussed earlier, the same end p.153

issues involved in the problem of the explanatory gap come out in the knowledge argument. When necessary I'll make explicit the connection between the two. Before presenting Perry's response to the knowledge argument, and in view of our discussion of the etiological constraint above, there is one preliminary point worth making. When considering Mary's new knowledge, it might be tempting to try to dismiss the problem immediately as follows. Look, one might say, of course Mary couldn't predict what it would be like to see red from what she knew of the physics and physiology of vision while in the Jackson room. After all, in order to possess the relevant concept—the phenomenal concept—with which to frame the relevant judgment concerning what it is like, one must first have the experience of seeing red oneself, and Mary hadn't had that yet. Her new knowledge upon emerging from the room is merely a matter of acquiring a new concept.

Though in fact the accounts we are going to investigate incorporate some aspects of this response, it's important to see from the outset that saying this much alone is clearly not sufficient. Nida-Rümelin (1995) has made the case quite convincingly. She asks us to consider Marianna, who starts out just like Mary, but instead of seeing a ripe tomato upon release, is ushered into a room with many abstract colored shapes hung on the walls, with no labels to say which color is which. She is asked if she can tell which of these is the color of the sky, of ripe tomatoes, and so on. It seems pretty clear that Marianna would not be in a position to know. When she is told that this one is red, she now learns what it is like to see red (or see ripe tomatoes). The knowledge seems new, despite the fact that she already possesses the first-person phenomenal concept of what it is like. Thus her inability to predict what it would be like, or to explain what it is like, given the conceptual resources available to her in the room is not merely a matter of her lacking the relevant experience. Coming to have the experience alone, and thereby acquiring the phenomenal concept, is not sufficient. The right sort of connection must also be made, and it is her inability to automatically establish that cognitive link between her previously acquired physical concepts and her newly acquired phenomenal concept that must be explained.

This is where the demonstrative account comes in. Perry asks us to consider the following situation. He has long admired Fred Dretske's work and knows it fairly well. In particular, he knows that Dretske is the author of *Knowledge and the Flow of Information*. However, he has never met Dretske and doesn't know what he looks like. At a party he meets a man he doesn't recognize and falls into a conversation about the topic of knowledge and information. He suggests to his conversational partner that he read *Knowledge and the Flow of Information*, proceeding then to present the main argument of the book. Much to his embarrassment, the man whom he's lecturing on the book informs him that he in fact wrote *Knowledge and the Flow of Information*.

So let's consider the following two statements:

1. Dretske wrote *Knowledge and the Flow of Information*.
2. This man wrote *Knowledge and the Flow of Information*.

Perry clearly knew (1) before meeting Dretske at the party. But (2) seems to be something he learned only after Dretske informed him of it in the middle of their end p.154

conversation. Before that moment, while conversing with Dretske (but not knowing his name), Perry presumably would have doubted (2) (otherwise Perry wouldn't have presumed to recommend that he read the book). So (2) seems to be a bit of new knowledge. Yet if we take "this man" to directly refer to Dretske, statements (1) and (2) express the same proposition, describe the same fact. Furthermore, though Perry doesn't himself say this, there doesn't seem to be any explanatory gap here. That is, we don't find ourselves puzzling about how this man could be Dretske.

The moral for the case of Mary is supposed to be straightforward. Let *Qr* stand for Mary's concept of the qualitative character the average human being experiences when seeing a ripe tomato in normal light. In the Jackson room, where Mary learns all the relevant physical and functional facts about visual experience, Mary learns in particular that

3. When Sally (a normal perceiver) sees a ripe tomato (in normal light), she occupies a state of type *Qr*.

After emerging from the Jackson room, while herself looking at a ripe tomato, Mary now comes to believe (she knows enough about her own brain and visual system to know that she herself is a normal perceiver):

4. When Sally sees a ripe tomato she occupies a state with *this* qualitative character (demonstrating the qualitative character of her own state).

Since, according to the materialist, the qualitative character she's picking out with her demonstrative is in fact *Qr*, (3) and (4) express the same proposition, describe the same fact. The problem with saying this is supposed to be that (4) seems to be a new belief, a new piece of knowledge. However, just as the epistemic novelty of (2) in no way impugns the identity of this man with Dretske, so too, argues Perry, the epistemic novelty of (4) in no way impugns the identity of this qualitative character with *Qr*. Furthermore, just as we have no problem understanding how this man could be Dretske, we should have no problem understanding how this qualitative character should be *Qr*.

Perry's analysis of what's going on in the party situation is instructive. He asks us to imagine that the mind is like a three-story building. On the first story are perceptual buffers, where files are opened for objects currently being perceived. Various features concerning the look of these objects are stored in these files. On the third floor are what he calls "detached" files, relatively permanent files for various objects that are not attached to any current perceptions. They store all sorts of information from all sorts of

sources, including memories of past perceptual encounters if there were any and information gleaned from reading and reports from others. The second story is dedicated to connecting files from stories one and three. Sockets hang down from story three, and plugs lead up from story one. When an object is recognized or when one learns the identity of an object currently perceived for the first time, the plug from the first-story file connects to the socket from the appropriate third-story file. So, in the party case, when Dretske told Perry who he was, the plug from the perceptual buffer file dubbed “this man” connected to the socket hanging down from the upper-story file named “Dretske,” and this
end p.155

allowed information to flow freely back and forth between the two files. Both files, of course, were about, or “of,” the same man all along.

As an armchair first approximation to what goes on in perceptual identification and recognition, this story seems pretty good. Let's suppose it's right. So when Perry finally learns that he's talking to Dretske himself—when the plug enters the socket—some genuinely new information does become available to Perry: namely, what Dretske looks like. “This man,” directly referential though it may be, does bring along with it the information that the object referred to is currently being perceived to have various sensible properties. That Dretske is currently being perceived by Perry to have these visual properties is genuinely new information, information not contained in the detached “Dretske” file he maintained up until now in his third-story file drawer. Though we do identify this man with Dretske, we don't identify any of the properties represented in the perceptual buffer file with any of the properties represented in the previously detached third-story file. Thus the sense that (2) expresses in some way a new piece of knowledge is easily explained by the fact that Perry now knows that the object presenting certain visual properties to him is the same object that wrote *Knowledge and the Flow of Information*. The object is the same, but not the properties associated with the two ways of picking it out.

But now, if we apply this model to Mary's case, the original problem reemerges. The idea is supposed to be that in formulating the thought expressed by (4), Mary is demonstrating her experience in much the same way that Perry is demonstrating Dretske in (2).

Presumably this is not supposed to be perception in the normal sense, but it is supposed to bear a strong resemblance to it. In particular, we are supposed to think of the new knowledge as a matter of linking plug and socket, as in the Dretske case. But there is a problem with this. Remember that *Qr* stands for Mary's previously detached concept of the relevant qualitative character. The file contains all sorts of neurophysiological, computational, and optical information. “This qualitative character” stands for the buffer file. But what does it contain? Well, it surely seems to contain some substantive idea of what it's like. The problem, then, is this. If we push the analogy with the Dretske case, then we can see easily enough how the two files can be “of” the same object—or, in this case, state. But an integral part of the Dretske case was that the information contained *within* the two files was different. If we maintain that aspect of the analogy in the Mary case, however, we end up reintroducing a new property, and we're back to where we started.

Well, maybe we're not supposed to take the analogy that literally. Maybe the feature of the Dretske case that involved the file folder containing perceptually derived information was inessential to the story. Perhaps we should think of the "buffer" in the Mary case as containing merely generic, topic-neutral information. That is, her first-person phenomenal representation has something like the form of "this sensory state," or "this qualitative character," so that the specific character is not itself represented in the file folder itself. In this way it would be quite different from the sort of file we're imagining in the standard perceptual buffer case. Or why not go further and think of the relevant file folder as just empty? Perhaps now we have a model of what's going on with Mary that explains her new "knowledge" in a benign way.

end p.156

A moment's reflection, however, reveals that the "empty file folder" idea won't work, for once we drop the assumption that the buffer file folder contains at least minimal information, we lose any grip on what's going on. Demonstratives, after all, are almost always associated with perceptual contents of some sort. I pick out something I see, hear, or touch with the expression, or thought, "*that F*" (where *F* represents some sortal or other), or even just "*that*." Even in the case where no sortal is employed, it's still clear that the demonstration is anchored to some perceptual representation. The whole idea of buffers containing files really only makes sense when there's something to put in the files. So although I can perfectly well make sense of the idea of demonstrating my current experience, we still need a model of what fills the role of the content material that anchors the demonstration: what's inside the buffer's file folder. The demonstrative itself cannot do all the work that's required here.

Demonstratives, then, seem to require some associated content material in order to lock onto their objects (whether they be objects or properties, particulars or universals). But there's still the option that that content material is quite thin and generic, just a matter of providing a sortal like "type of qualitative character" or "type of sensory state." So the idea is this. When Mary entertains (4), she is identifying the same property, or type of state, that is picked out by her rich scientific description with the one picked out by her demonstration of "this state," where the only substantive characterization available in the demonstrative file is that it is a state she is in. No qualitative or descriptive content is associated with this way of picking it out.

Returning to the Dretske case, the better analogy would be the following scenario. I've met Fred Dretske before, know perfectly well what he looks like, sounds like, and so on. However, it turns out that he has a twin brother who looks exactly like him, talks like him, and holds the same views on knowledge and information. In fact, his name is "Fred" as well. Of course, Fred2 (as I'll call him, not what he calls himself) didn't write *Knowledge and the Flow of Information*. So when speaking with someone whom I know to be one of the Freds at a party, I might wonder which one it is. Then, after Fred tells me he wrote *Knowledge and the Flow of Information*, I will have learned something new, which I can express as (2).

This situation is now quite similar to the one we're imagining Mary is in when she expresses (4). It's not quite the same, since, as in Perry's scenario, it is still the case that the perceptual buffer's file folder contains quite a bit of information concerning the look

and sound of the man I'm speaking to. However, because of my previously established familiarity with Fred, none of this information (or, at least, none of it that remains with me, or achieves salience) contained within the buffer's file is new. The file's plug connecting to the socket from above achieves no information flow over and above the mere fact that *this* is the one that *that* upper-story file is about (as opposed to the one I have under the name "Fred2").

However, I think it's clear that this modified Dretske scenario is not a good model for what's going on in the Mary case, precisely because, on this scenario, no real new information is introduced via the demonstrative presentation. Mary doesn't seem to learn just that the state she can describe in such rich theoretical vocabulary is happening here and now; that it's *this* one. She forms a new

end p.157

conception, one with substantive and determinate content, of this state. The situation is much more like Perry's original Dretske case, where he learns what the author of *Knowledge and the Flow of Information* looks like. That is, in the Mary case, just as in Perry's version of the Dretske case, the new information is not constituted only by the mere linking of buffer file with detached file, but also by the fact that there is something new in that buffer file itself: a new way of representing this particular type of qualitative character. It is to this new representation, contained within the buffer file, that we must look for the phenomenal concept we're after.

Before turning to our next option, I want to return to a thread I earlier left hanging. Recall that I described a materialist who accepted the view that most ordinary nonphenomenal concepts were analyzable in such a way as to permit upward derivations from the relevant microlevel descriptions, but who maintained that phenomenal concepts were primitives and therefore were not subject to upward derivations. On this view, it was the presence of upward derivability in the standard nonphenomenal cases and the absence of it in the phenomenal case that explained the core contrast. However, our investigation of the demonstrative model shows that mere absence of upward derivability cannot be the whole story. Descriptions that essentially involve indexicals or demonstratives cannot be derived from descriptions lacking them. Yet, as we see from the modified Dretske case, no sense of substantively new information, or a distinct property, is automatically engendered. There must be something about the way that phenomenal concepts afford a grasp of their objects—afford a kind of “cognitive presence” to phenomenal properties—that explains why they seem distinct from anything conceived by another method. Providing a model of this relation, of what this cognitive presence amounts to, that accords with the Materialist Constraint is the challenge.

I used the term “cognitive presence” to try to express the unique relation that phenomenal concepts seem to bear to what they represent, but I might just as easily have used the traditional Russellian term “acquaintance.” Russell (1912, chap. 5) divided the objects of knowledge and thought into those that were known by acquaintance and those that were known by description. Among the objects knowable by acquaintance were the immediate contents of sensory experience. Most ordinary objects were known only by description, where the descriptions in question contained logical terms and those representing items with which we were acquainted. Thus for Russell, all epistemic access and reference bottomed out with acquaintance.

Let's put aside the foundationalist epistemology and theory of reference, allowing that terms representing items with which we are not acquainted might pick out their objects, and afford epistemic access to them, without employing modes of presentation that are ultimately constructed from objects with which we are acquainted. One might still find the distinction between items (whether they be objects or properties) with which we are acquainted and those with which we are not acquainted useful; though, if we are eschewing Russell's foundationalism, the latter category shouldn't be characterized as those items we know "by description." Let's allow, instead, that there are two forms of direct reference: one involving acquaintance and one not.

end p.158

The essential idea behind Russell's distinction still remains. When it comes to the properties of our immediate experience, we stand in a kind of epistemic relation to them that is more intimate, more substantive, than the kind of relation that obtains between our minds and other items. The properties of experience are, to use my other phrase, cognitively present to us. The idea, then, is to explain why it so strongly seems to be the case that what is presented by way of phenomenal concepts is distinct from what is presented by nonphenomenal concepts by appeal to the distinction between acquaintance and other forms of representation. Because phenomenal concepts afford acquaintance, whereas nonphenomenal concepts do not, even if they in fact pick out the very same properties, we find it cognitively difficult to see how this can be.

Perhaps we can put it this way. Nonphenomenal concepts either pick out their objects by way of substantive modes of presentation constructed out of other concepts, or by direct labeling that involves no substantive mode of presentation at all—merely, say, a causal relation of some sort between the concept and what it's a concept of. By contrast, phenomenal concepts employ a substantive mode of presentation and use what they are about—phenomenal properties—as the modes of presentation.⁴ Since phenomenal properties are themselves involved in the very mode of presentation when conceived via phenomenal concepts, but not when conceived via nonphenomenal concepts, the referents of the two sorts of concepts will present themselves to us as distinct, even though they are identical.

I realize that the notions of acquaintance, of cognitive presence, and of phenomenal properties being their own modes of presentation are all still too metaphorical. But let's suppose we have enough of a handle on them to proceed. What we're looking for is some model of this special cognitive relation that supposedly obtains between a phenomenal concept and its corresponding phenomenal property that simultaneously satisfies the Materialist Constraint. This means that whatever acquaintance is, it can't be a basic relation; it must be constructible out of other, nonmental relations.

In this section, I want to investigate the possibilities for essentially providing a materialist model of acquaintance. There are two approaches I want to look at, and they share a crucial feature: both try to incorporate an instance of the phenomenal property into the phenomenal concept itself. In this way, they attempt to capture the idea that the phenomenal property (or an instance of it) serves as its own mode of presentation, which we can for now take to be the essential element in the relation of acquaintance.

The first approach is really a modification of the demonstrative model. In standard cases of demonstrative representations, there are three elements in play: the demonstrative component itself, a perceptual representation, and the object demonstrated. Notice that the demonstrative representation, or concept, itself contains only the first two elements mentioned. The object is not itself part of the
end p.159

representation; it is merely what is represented. So, for instance, in Perry's Dretske case, his thought "that man" picks out Dretske by way of demonstrating whatever it is that presents the particular appearance he is currently perceptually aware of. In terms of his file model, the perceptual information is contained in the file, while the file itself plays the role of the demonstrative element. "That man" refers to whomever this perceptual information is about.

But suppose that, instead of treating the demonstrative representation as composed of the first two elements, which together pick out the third, we collapse the second two elements into one. That is, the demonstrative representation in this case is taken to consist of the demonstrative element together with what is demonstrated, the phenomenal state (or, to be more precise, the property it is currently instantiating). If we like, in keeping with Perry's metaphorical architecture, we can picture this as putting the phenomenal state itself into the demonstrative file folder. A phenomenal concept, then, is a complex state consisting of a demonstrative together with the state demonstrated, interpreted to represent the relevant property instantiated by the demonstrated state. To be acquainted with a property is to demonstrate an instance of it in a state that includes the instance as a component of the representation.

The second approach is to forget about the demonstrative element altogether, and just let tokens of phenomenal states themselves serve as representations of the phenomenal properties they instantiate. In other words, phenomenal concepts are tokened by the very same states that serve to instantiate the properties they represent. Again, this is a very graphic way to implement the notion that a phenomenal property serves as its own mode of presentation.

As a first step in evaluating these two approaches, I want to argue that in fact there really is no significant difference between them; they amount to the same model in the end. The argument takes the form of showing how each version presupposes, or incorporates, the other one. Let's start with how the demonstrative version incorporates the self-representation version. Notice that the idea of putting the phenomenal state itself—the demonstratum (or the current instantiation of the demonstratum)—into the demonstrative file itself, as playing the role of mode of presentation, is really quite odd. In the standard file story, there are two ways of looking at the contents of the file. For instance, in Perry's original Dretske case, the demonstrative buffer file contains the perceptual information concerning the appearance of the man with whom Perry is speaking. This perceptual information, however, can be thought of as a physical token of some sort—imagine that it is literally a picture stored in a file—or it can be thought of as a representation of an appearance. Clearly in Perry's story about meeting Dretske, it is the second way of thinking about the contents of the file that is relevant. The intrinsic features of the vehicle are irrelevant.

The point is that the contents of the file, which serve as the mode of presentation, bring us into contact with the object of the file by representing certain of its properties. It is a total distortion of the "file" model to stick the object of the file itself into the file and say that it serves as its own mode of presentation, unless, of course, we are already taking for granted that it is representing itself. But then we might as well do away with the file altogether. The entire burden of representing the

end p.160

phenomenal property is borne by the phenomenal state itself. To let the object itself serve as the mode of presentation of the demonstrative representation that picks it out is already to presume the kind of self-representation involved in the second model.

Now let's take it in the other direction. The idea is supposed to be that when a phenomenal property is instantiated by a token phenomenal state, it is thereby represented by this token. The question that immediately arises, of course, is how the representation relation is established in this case. After all, normally, instantiating a property is neither necessary nor sufficient for representing it. So what is it in this case that makes the instantiation of the property also a representation of it?

Clearly, the answer must lie with the functional role occupied by phenomenal states.

Most representations represent what they do by virtue of some causal/ informational link with their referents. In some cases, however, semantic significance can arise purely as a matter of functional role. A good example of this might be the logical constants. It's hard to see how a symbol meaning conjunction could acquire that meaning by virtue of some causal/informational link with the relevant truth-function: what would that mean?

However, one could see how a symbol might count as representing conjunction by virtue of its interactions with other symbols. Using this analogy, one might imagine that by virtue of some particular pattern of causal interactions with other representational states, phenomenal states could acquire interpretations that involved reference to the very phenomenal properties they instantiate.

Let's assume, then, that the representation of phenomenal properties is effected in some way by functional role. There are two ways to interpret this appeal to functional role. The first is to imagine that for each phenomenal property, there is a unique functional role that serves to designate it (or, to be more precise, is such that by virtue of playing that role, a state represents it). There are two reasons this couldn't be right. First, there are just too many distinguishable phenomenal properties for there to be a distinct type of functional role corresponding to each. It's also unclear how a functional role could even serve this function of picking out a particular phenomenal property. But even if this problem could be overcome, this way of interpreting the appeal to functional role doesn't really capture the whole idea of using a phenomenal state as its own mode of presentation, of realizing the relation of acquaintance. After all, if playing functional role *Fr* is what makes a state represent phenomenal property *R*, then the fact that it's an instance of *R* that is playing the role representing *R* would be beside the point. It could just as easily have been some other state. What's needed, then, is a way of interpreting the appeal to functional role that takes directly into account the identity of the role filler.

The second way of interpreting the appeal to functional role is to imagine a single type of role covering all phenomenal states/properties, one that serves to pick out whatever

phenomenal property it is that is instantiated by the token filling the role. This would make the identity of the token phenomenal state an essential part of the representation. In this case, we would have a genuine case of self-representation. Phenomenal states, by virtue of playing this particular functional

end p.161

role, would be interpreted as “saying” something like “the phenomenal property I am currently instantiating” or just “this phenomenal property.”⁵

At this point it should be clear why the self-representation model really amounts to the same thing as the demonstrative model. The functional role by virtue of which the phenomenal state is representing the phenomenal property it instantiates serves as a demonstrative, or indexical. How, after all, would one implement such a role? One obvious way would be to use location. Symbols occupying certain locations would be interpreted as referring to themselves, or something like that. But then the location is really just like a pointer; it has the same significance as a symbol saying “this one.”⁶ The crucial feature of the model under consideration—whether or not an explicitly demonstrative element is present—is the presence of the phenomenal state itself within the implementation of the corresponding phenomenal concept. The physical presence of an instance of the phenomenal property is thereby supposed to explain the especially intimate cognitive relation afforded by phenomenal concepts. In other words, a phenomenal concept affords acquaintance with the relevant phenomenal property by containing an instance of that property within it.

It's hard to imagine how else a physicalist could capture cognitive immediacy, or acquaintance. If physically placing the relevant state right into the structure that realizes a phenomenal concept doesn't do it, what could? Yet, it seems to me that putting the matter this starkly merely highlights the model's inadequacy. Acquaintance, or cognitive presence, or whatever it is that is supposed to constitute the especially immediate and intimate cognitive relation between phenomenal concepts and their objects, is just that: a *cognitive* relation. It is not at all clear why, or how, *physical* presence translates into cognitive presence. In general, when considering the cognitive properties of a representational system, the physical identities of the implementing tokens are irrelevant. What matters is how the various tokens relate to each other. Their relations to their objects matter only to the extent that it is necessary to determine from those relations what they represent. But once that is determined, it is unclear how differences in the mechanisms of the representation relation are supposed to explain differences in cognitive significance.

We must remember here what we're after. The bottom line is that we want an explanation of the existence of the explanatory gap that satisfies the Materialist Constraint. Why should the phenomenal character of a sensory experience seem inexplicable in terms of the corresponding physiological states? In particular, how could there even be a question of explaining the phenomenal character if it *just is* a certain physiological property? Furthermore, as we saw earlier, when considering creatures physically different from us, we can't help feeling that there is a

end p.162

substantive, nonsemantic open question regarding the nature and/or existence of their conscious experience. But this must be an illusion if the materialist identity theory is correct. Clearly something about the way phenomenal states present themselves to us in the first-person mode make them seem distinct from whatever is represented by their third-person descriptions. This is the source of the idea that the explanatory gap arises here, in the psychophysical case, but not in other cases of theoretical reductions, because of the peculiar concepts by which we represent phenomenal properties in standard first-person access.

After surveying various candidates for the peculiar feature on which to pin the blame, we saw that it isn't merely a matter of representational or epistemic primitiveness. The problem isn't that the conception of phenomenal properties afforded by phenomenal concepts is too thin, lacking connections to other concepts. If anything, the problem seems to be quite the opposite. The first-person access we have to the properties of experience seems quite rich; we are afforded a very substantive and determinate conception of a reddish experience merely by having it. The idea, then, is that it makes sense that we would find it hard to see how that with which we were acquainted in this way could be the very same thing as that which is picked out without the benefit of acquaintance.⁷ But if this is what we're after, then it doesn't help merely to find some functional difference between phenomenal concepts and others. We need to find a difference that plausibly reconstructs the acquaintance/non-acquaintance distinction. The proposal under consideration is supposed to do just that. Acquaintance is explained by the physical presence of the represented within the representation itself. But, I ask again: How is physical presence an explanation of cognitive presence? Sure, when we think of a relation such as acquaintance, with its sense of immediacy, metaphorical language about "sticking the object right in there" irresistibly comes to mind. But this language is metaphorical. The current proposal is to solve the problem by taking it literally.

Perhaps, in the end, this is the right way to go. But we are still owed an account of how physical presence alone is responsible for cognitive presence. That is, how does the presence of the relevant state within the physical implementation of the representation become something of which we are aware? It still smacks of that famous cartoon of the physicist scribbling all these formulas on the blackboard with this one circle in the middle that says "and then a miracle occurs." The transition from physical containment to awareness—the special kind allegedly afforded by phenomenal concepts—is still an inexplicable transition. It is subject to its own explanatory gap, just as much as is the original relation between phenomenal properties and their physical correlates.⁸
end p.163

There are two points I want to address in conclusion. First, someone might object to my entire argument as follows.⁹ What you are asking for, goes the objection, is, in effect, to bridge the explanatory gap. But this is precisely what we materialists admit can't be done, as one could predict from the various models surveyed here. So you are asking for the impossible, and the inability to do the impossible does not count against any theory. My response is that I'm only requiring of the various models surveyed that they accomplish what they are advertised to accomplish: namely, explain why there is an

explanatory gap. Granted, I'm asking for an explanation. But remember the original gap separated phenomenal properties from their underlying physical mechanisms. This gap I am not asking to be bridged. However, materialists claim that though they cannot explain phenomenal properties in terms of physical properties, they *can* explain why they can't explain it. The various models of how to
end p.164

implement phenomenal concepts is supposed to accomplish precisely that explanatory task. So, since the challenge that materialists say they can meet is itself to provide an explanation—in this case, of the fact that there is an explanatory gap between the physical and the phenomenal—it's a perfectly legitimate objection to point out that they haven't provided the requisite explanation. One might say that there now is a second explanatory gap: between implementations of cognitive architecture and whatever it is about phenomenal concepts—in my terms, that they afford genuine cognitive presence to phenomenal properties—that is responsible for the original explanatory gap. If one thought the original explanatory gap was a problem and needed to be explained away, then one ought to be bothered by this one as well.

Notice that nothing I've argued is intended to show that phenomenal properties aren't physical properties in the end. For that matter, phenomenal concepts may indeed be physically realized in one of the ways described above. The problem is that we don't understand how either story could be true, how the features we encounter in experience, or the encountering relation itself, could turn out to be a neural mechanism. This is indeed a situation materialists should find troubling.

Finally, I want to end with a bit of speculation. Suppose I'm right that we can't now imagine how a materialist story of phenomenal concepts would go. No mere physical-causal mechanism can provide the kind of cognitive presence we seem to enjoy with respect to our phenomenal experience. So what is it we need? It seems to me that we need something like the old-fashioned relation of acquaintance. We are acquainted with our experience, and as acquaintance *presents* properties, not merely represents them, we find it difficult to integrate what is presented with what is only represented in a way that allows the latter to explain the former. If acquaintance itself cannot be explained in terms of physical-causal mechanisms, as I claim (at least so far) it can't, then we have to contemplate the possibility that it is a brute relation. If so, then the Materialist Constraint is violated, and materialism is false.

It could turn out, then, that materialism is false not because phenomenal properties themselves are not physical—they may yet be for all we know. Rather, it would be false, on this view, because somehow we embody a relation to them that is itself brute and irreducible to physical relations. Is this a coherent position? Could phenomenal properties be physical while acquaintance is not? I don't know, but the question, to my mind, deserves exploration.

But haven't we already done away with acquaintance, sense data, and the “myth of the given”? One answer is to just say, “maybe not.” On the other hand, nothing in what I've said entails that there are sense data or that acquaintance plays the same foundational epistemic role assigned to it by Russell and the positivists. However, I do want to acknowledge that there is much to puzzle about in allowing a relation of acquaintance.

Does acquaintance entail “revelation,” the doctrine that the essential nature of that with which we are acquainted is revealed thereby? Indeed, is the notion even coherent in the end? That is, if we abandon materialism for its inability to explain phenomenal experience, do we then flirt with outright incoherence instead? These, too, are questions to which I have no answer at present. But, again, I think they deserve exploration.

end p.165

Acknowledgments

Earlier versions of this chapter were presented to the NEH Institute on Consciousness and Intentionality, University of California—Santa Cruz (2002); the Jowett Society at Oxford University; the University of St. Andrews; Ohio State University; Bowling Green University; the conference on “Consciousness: Conceptual and Explanatory Issues” in Magdeburg, Germany; the University of Canterbury, Christchurch, New Zealand; the Australian National University; and the University of Otago, Dunedin, New Zealand. I want to thank all the members of these audiences for their helpful comments and criticisms. I especially want to thank Katalin Balog, Janet Levin, and Susanna Siegel for their comments on earlier drafts.

References

- Balog, K. (2002). The “Quotational Account” of Phenomenal Concepts. Unpublished.
- Hill, C. S., and McLaughlin, B. P. (1999). There Are Fewer Things in Reality Than Are Dreamt of in Chalmers' Philosophy. *Philosophy and Phenomenological Research* 59: 445–54. [Link](#)
- Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36. [Link](#)
- Loar, B. (1990/97). Phenomenal States. In *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif: Ridgeview. Revised version in *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzeldere: 597–616. Cambridge: MIT Press, 1997.
- Nida-Rümelin, M. (1995). What Mary Couldn't Know: Belief about Phenomenal States. In *Conscious Experience*, ed. T. Metzinger: 219–42. Paderborn: Schöningh/ Imprint Academic.
- Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press. [Link](#) [OSO X-Reference](#)
- Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press.
- Russell, B. (1912). *The Problems of Philosophy*. New York: Oxford University Press.
- Sturgeon, S. (2000). *Matters of Mind*. London: Routledge.
- Tye, M. (2000). *Consciousness, Color, and Content*. Cambridge: MIT Press.
- end p.166

nine Phenomenal Concepts and the Explanatory Gap

Confronted with the apparent explanatory gap between physical processes and consciousness, philosophers have reacted in many different ways. Some deny that any explanatory gap exists at all. Some hold that there is an explanatory gap for now, but that it will eventually be closed. Some hold that the explanatory gap corresponds to an ontological gap in nature.

In this chapter, I want to explore another reaction to the explanatory gap. Those who react in this way agree that there is an explanatory gap, but they hold that it stems from the way we *think* about consciousness. In particular, this view locates the gap in the relationship between our concepts of physical processes and our concepts of consciousness, rather than in the relationship between physical processes and consciousness themselves.

Following Stoljar (2005), we can call this the *phenomenal concept strategy*. Proponents of this strategy argue that phenomenal concepts—our concepts of conscious states—have a certain special nature. Proponents suggest that given this special nature, it is predictable that we will find an explanatory gap between physical processes conceived under physical concepts, and conscious states conceived under phenomenal concepts. At the same time, they argue that our possession of concepts with this special nature can itself be explained in physical terms.

If this is right, then we may not have a straightforward physical explanation of consciousness, but we have the next best thing: a physical explanation of why we find an explanatory gap. From here, proponents infer that the existence of the explanatory gap is entirely compatible with the truth of physicalism. From there, they infer that there can be no sound argument from the existence of the explanatory gap to the falsity of physicalism.

In addition, proponents often use this strategy to deflate other intuitions that lead some to reject physicalism about consciousness: intuitions about conceivability and about knowledge, for example. They suggest that these intuitions are consequences of the special nature of phenomenal concepts (which, again, can itself be explained
end p.167

in physical terms). They conclude that these intuitions cannot give us conclusive reason to reject physicalism.

This extremely interesting strategy is perhaps the most attractive option for a physicalist to take in responding to the problem of consciousness. If it succeeded, the strategy would respect both the reality of consciousness and the epistemic intuitions that generate the puzzle of consciousness while explaining why these phenomena are entirely compatible with physicalism.

I think that the strategy cannot succeed. On close examination, we can see that no account of phenomenal concepts is both powerful enough to explain our epistemic situation with regard to consciousness and tame enough to be explained in physical terms. That is, if the relevant features of phenomenal concepts can be explained in physical terms, the features cannot explain the explanatory gap. And if the features can explain the

explanatory gap, they cannot themselves be explained in physical terms. In what follows I will explain why.

Epistemic Gaps and Ontological Gaps

Let P be the complete microphysical truth about the universe: a long conjunctive sentence detailing the fundamental microphysical properties of every fundamental microphysical entity across space and time. Let Q be an arbitrary truth about phenomenal consciousness: for example, the truth that somebody is phenomenally conscious (that is, that there is something it is like to be that person) or that I am experiencing a certain shade of phenomenal blueness.

Many puzzles of consciousness start from the observation that there is an apparent *epistemic gap* between P and Q : a gap between knowledge of P and knowledge of Q , or between our conception of P and our conception of Q .

Take Frank Jackson's case of Mary in the black-and-white room, who knows all the microphysical facts but who still does not know what it is like to see red. It appears that Mary may know P and may have no limitations on powers of a priori reasoning, but may still fail to know Q (where here Q is a truth about what it is like for ordinary people to see red things). This suggests that the truth of Q is not deducible by a priori reasoning from the truth of P . More specifically, it suggests that the material conditional $P \supset Q$ is not knowable a priori.

Or take the conceivability of zombies. A zombie is a hypothetical creature that is physically identical to a conscious being but is not conscious at all. Many people hold that zombies are conceivable in principle, and they hold further that in principle one could conceive of a zombie world: one that is physically identical to ours, but without consciousness. Many people also hold that we can conceive of an *inverted world*: one that is physically identical to ours, but in which some conscious states differ from the corresponding states in our world. If this is right, then there is a gap between conceiving of P and conceiving of Q . It appears that $P \& \sim Q$ is conceivable, where Q is a truth such as "Someone is phenomenally conscious" (in the first case), or a truth specifying a particular state of phenomenal consciousness (in the second).

(I will not say much about exactly what conceivability involves because most of what I say will be compatible with various understandings of conceivability.
end p.168

But at minimum, we can say that the conceivability of S requires that the truth of S cannot be ruled out a priori. This is the notion that I have elsewhere called *negative* conceivability [strictly: ideal primary negative conceivability]. One may also suggest that the conceivability of S requires that one can clearly and distinctly imagine a situation in which S is the case. This is the notion that I have elsewhere called *positive* conceivability [strictly: ideal primary positive conceivability]. I think that positive conceivability is the canonical notion of conceivability, but for the most part, the arguments in this chapter can

operate with either notion. In those cases in which the distinction is relevant, I will make it explicit. For much more on these notions of conceivability, see Chalmers 2002.)

Many hold further that these epistemic gaps go along with an explanatory gap between P and Q . The explanatory gap comes from considering the question, Why, given that P is the case, is Q the case? (Why, given that P is the case, is there phenomenal consciousness? And why are there the specific conscious states that there are?) The gap is grounded in part in the apparent inability to deduce Q from P : if one cannot deduce that Q is the case from the information that P is the case, then it is hard to see how one could explain the truth of Q *wholly* in terms of the truth of P . It is grounded even more strongly in the conceivability of P without Q . If one can conceive of a world that is physically just like this one but without consciousness, then it seems that one has to add something more to P to explain why there is consciousness in our world. And if one can conceive of a world that is physically just like this one but with different states of consciousness, then it seems that one has to add something more to P to explain why conscious states are the way they are in our world.

From these epistemic gaps, some infer an ontological gap. One may infer this ontological gap directly from the explanatory gap: if we cannot explain consciousness in terms of physical processes, then consciousness cannot be a physical process. Or one may infer it from one of the other epistemic gaps. For example, one may infer from the claim that $P \& \sim Q$ is conceivable that $P \& \sim Q$ is metaphysically possible, and conclude that physicalism is false. If there is a possible world physically just like this one but without consciousness, then the existence of consciousness is an ontologically further fact about our world.

At this point, materialists typically respond in one of two ways. Type-A materialists deny the epistemic gap. Paradigmatic type-A materialists deny there is any factual knowledge that Mary lacks inside her black-and-white room; they deny that zombies are conceivable, at least on ideal reflection; and they deny that there is an explanatory gap that survives reflection. Type-A materialism is an important view. But proponents of the phenomenal concept strategy reject type-A materialism, so I will not discuss it further here.

Type-B materialists accept that there is an epistemic gap but deny the inference to an ontological gap. Paradigmatic type-B materialists hold that Mary lacks knowledge, but not of ontologically distinct facts about the world; they hold that zombies are conceivable but not metaphysically possible; and that although there may be no satisfying explanation of consciousness in physical terms, consciousness is a physical process all the same.
end p.169

Type-B materialists typically embrace *conceptual dualism* combined with *ontological monism*. They hold that phenomenal *concepts* are distinct from any physical or functional concepts. But they hold that phenomenal *properties* are identical to certain physical or functional properties, or at least that they are constituted by these properties in such a way that they supervene on them with metaphysical necessity. In this view, conceptual dualism gives rise to the explanatory gap, whereas ontological monism avoids any ontological gap.

Here type-B materialists often appeal to analogies with other cases in which distinct concepts refer to the same property. “Heat” and “molecular motion” express distinct concepts, for example, but many hold that they refer to the same property. By analogy, some type-B materialists suggest that a phenomenal term (e.g., “pain”) and a physical term (e.g., “C-fiber firing”) may express distinct concepts but pick out the same property. More generally, type-B materialists typically hold that the material conditional “ $P \supset Q$ ” is an instance of Kripke's necessary a posteriori: like “water is H_2O ,” the conditional is not knowable a priori, but it is true in all possible worlds. If successful, these analogies would reconcile the epistemic gap with ontological monism.

However, the success of these analogies is widely disputed. Kripke himself argued that the relation between mental and physical expressions is different in kind from the relation between “heat” and “the motion of molecules,” or that between “water” and “ H_2O ,” so that the grounds for a posteriori identities or necessities in these standard cases are not present in the mental-physical case. Since then, many opponents and even proponents of type-B materialism have argued that mental and physical properties are not analogous. Some (e.g., White 1986, Loar 1990/97) argue that in the standard cases, the distinct concepts (e.g., “heat” and “the motion of molecules”) are associated with distinct properties at least as modes of presentation of their referent, if not as their actual referent. Some (e.g., Chalmers 1996, 2002) argue that the standard cases are all compatible with an attenuated link between conceivability and possibility, expressible using two-dimensional semantics. Some (e.g., Jackson 1998) argue that the standard cases are all compatible with the thesis that physicalism requires a priori entailment of all truths by physical truths. Some (e.g., Levine 2001) argue that the physical-phenomenal case involves a “thick” explanatory gap that is unlike those present in the standard cases. These differences strongly suggest that the standard way of reconciling conceptual dualism with ontological monism does not apply to the conceptual dualism between the physical and the phenomenal. If the principles that hold in the standard cases applied here, then the conceptual dualism would lead to an ontological dualism. For example, we would expect distinct properties to serve as modes of presentation for physical and phenomenal concepts; and from here one can reason to an underlying ontological dualism at the level of these properties. Likewise, we would expect there to be some metaphysically possible world in the vicinity of what we conceive when we conceive of zombies and inverts; and from here one can reason to a failure of metaphysical supervenience of everything on the physical. If so, then the epistemic gap will once again lead to an ontological gap.

end p.170

The Phenomenal Concept Strategy

Partly to avoid these problems, many type-B materialists have turned to a different strategy for reconciling conceptual dualism and ontological monism. Instead of focusing on quite general features of a posteriori identities and necessities, this strategy focuses on features that are specific to phenomenal concepts. Proponents of the phenomenal concept strategy typically allow that we are faced with a distinctive epistemic gap in the physical-

phenomenal case, one that is in certain respects unlike the epistemic gaps that one finds in the standard cases. But they hold that this distinctive epistemic gap can be explained in terms of certain distinctive features of phenomenal concepts. And they hold that these distinctive features are themselves compatible with an underlying ontological monism.

Recognitional concepts: The *locus classicus* for the phenomenal concept strategy is Brian Loar's paper "Phenomenal States" (1990/97), in which he suggests that phenomenal concepts are recognitional concepts that pick out their objects via noncontingent modes of presentation. (Related proposals involving recognitional concepts are made by Carruthers [2004], Tye [2003], and Levin [chap. 6, this volume].) Recognitional concepts are concepts deployed when we recognize an object as being one of *those*, without relying on theoretical knowledge or other background knowledge. For example, we may have a recognitional concept of a certain sort of cactus. One may also have a theoretical concept of that sort of cactus, so that there are two concepts referring to the same sort of entity. In standard cases, these two concepts will be associated with distinct properties as modes of presentation (for example, one's recognitional concept of a cactus may be associated with the property *typically causes such-and-such experience*), so this will not ground a full-scale ontological monism. But Loar suggests that phenomenal concepts are special recognitional concepts because the property that is the referent also serves as a mode of presentation. He argues that this special character of phenomenal concepts explains the distinctive epistemic gap in a manner that is compatible with ontological monism.

Distinct conceptual roles: Developing a suggestion by Nagel (1974), Hill (1997; see also Hill and McLaughlin 1999) suggests that phenomenal concepts and physical concepts are associated with distinct faculties and modes of reasoning, and that they play very different conceptual roles. Hill argues that the distinctive epistemic gaps between the physical and phenomenal are explained by this distinctness in conceptual roles, and he suggests that we should expect the epistemic gaps to be present even if the distinct concepts refer to the same property.

Indexical concepts: A number of philosophers (including Ismael [1999], O'Dea [2002], and Perry [2001]) have suggested that phenomenal concepts are a sort of indexical concept, analogous to *I* and *now*. There are familiar epistemic gaps between objective and indexical concepts, noted by Perry (1977) and many others. For example, even given complete objective knowledge of the world, one might not be able to know what time it is now, or where one is located. Proponents of the indexical concept strategy suggest that the epistemic gap between the physical and phenomenal has a similar character. On this view, just as "now" picks out a certain

end p.171

objective time under an indexical mode of presentation, phenomenal concepts pick out states of the brain under an indexical mode of presentation.

Quotational concepts: Finally, some philosophers have suggested that phenomenal concepts are special because their referents—phenomenal states—serve as constituents of the concepts themselves (or as constituents of the corresponding mental representations). Sometimes this view of phenomenal concepts is put forward without any associated ambition to support type-B materialism (e.g., Chalmers 2003a). But some, such as

Papineau (2002 and chap. 7, this volume) and Block (chap. 12, this volume), suggest that this view of phenomenal concepts can explain the epistemic gap in terms acceptable to a materialist. For example, Papineau sees phenomenal concepts as *quotational concepts*, which represent their referent as *That state: —*, where the blank space is filled by an embedded phenomenal state in a way loosely analogous to the way that a word might be embedded between quotation marks. Papineau suggests that even if the embedded state is a neural state, this quotational structure will still give rise to the familiar epistemic gaps. Other proponents of the phenomenal concept strategy include Sturgeon (1994), who proposes that the explanatory gap is grounded in the fact that phenomenal states serve as their own canonical evidence; Levine (2001), who suggests that phenomenal concepts may crucially involve a nonascriptive mode of presentation of their referent; and Aydede and Güzeldere (2005), who give an information-theoretic analysis of the special relation between phenomenal concepts and perceptual concepts.

I have discussed many of these views elsewhere. (See Chalmers 1999 for discussion of the first two views, and Chalmers 2003a for discussion of the third and fourth.) Here I will focus instead on what is common to all the views, arguing on quite general grounds that no instance of the phenomenal concept strategy can succeed in grounding a type-B materialist view of the phenomenal. Later I will apply this general argument to some specific views.

The general structure of the phenomenal concept strategy can be represented as follows. Proponents put forward a thesis *C* attributing certain psychological features—call these the key features—to human beings. They argue (1) that *C* is true: humans actually have the key features; (2) that *C* explains our epistemic situation with regard to consciousness: *C* explains why we are confronted with the relevant distinctive epistemic gaps; and (3) that *C* itself can be explained in physical terms: one can (at least in principle) give a materialistically acceptable explanation of how it is that humans have the key features. This is a powerful strategy. If it is successful, we may not have a direct physical explanation of consciousness, but we will have the next best thing: a physical explanation of the explanatory gap. One might plausibly hold that if we have a physical explanation of all the epistemic data that generate arguments for dualism, then the force of these arguments will be undercut. I think this matter is not completely obvious—one might hold that the residual first-order explanatory gap still poses a problem for physicalism—but I will concede the point for the purposes of this chapter. There is no question that a physical explanation of the relevant epistemic gaps would at least carry considerable force in favor of physicalism.

end p.172

Note that for the strategy to work, all three components are essential. If (1) or (2) fail, then the presence of the relevant epistemic gaps in us will not be explained. If (3) fails, on the other hand, then although thesis *C* may help us understand the conceptual structure of the epistemic gap, it will carry no weight in deflating the gap. If the epistemic gap is grounded in special features of phenomenal concepts that are not physically explainable, then these features will generate a gap of their own. Opponents of the strategy will then argue that the special features themselves require a nonphysical explanation, and may plausibly suggest that the special features themselves reflect the presence of irreducible

phenomenal experience. If so, the phenomenal concept strategy will do little to support physicalism.

It should be noted that not all proponents of the phenomenal concept strategy are explicitly committed to (3), the thesis that the relevant features of phenomenal concepts must be physically explicable. Some proponents, such as Loar and Sturgeon, are silent on the matter. Almost all of them, however, use the phenomenal concept strategy to resist the inference from the epistemic gap to an ontological gap. I will argue later that without (3), the phenomenal concept strategy has no force in resisting this inference.

There is a related strategy that I will not discuss here. This is the type-A materialist strategy of appealing to psychological features to explain why we have false beliefs or mistaken epistemic intuitions about consciousness (see, for example, Dennett 1981 and chap. 1, this volume; and Jackson 2003 and chap. 3, this volume). In its most extreme form, this strategy may involve an attempted psychological explanation of why we think we are conscious, when in fact we are not. In a less extreme form, the strategy may involve an attempted psychological explanation of why we think there is an epistemic gap between physical and phenomenal truths when in fact there is not. For example, it may attempt to explain why we think Mary gains new knowledge when in fact she does not, or why we think zombies are conceivable when in fact they are not. This is an important and interesting strategy, but it is not my target here. My target is, rather, a type-B materialist who accepts that we are phenomenally conscious and that there is an epistemic gap between physical and phenomenal truths, and who aims to give a psychological explanation of the existence of this epistemic gap.

A Master Argument

I will argue that no account can simultaneously satisfy (2) and (3). For any candidate thesis *C* about psychological features of human beings, then either

1. *C* is not physically explicable

or

2. *C* does not explain our epistemic situation with regard to consciousness.

Here the key question will be: is $P \& \sim C$ conceivable? That is, can we conceive of beings physically identical to us (in physically identical environments, if necessary) that do *not* have the psychological features attributed by thesis *C*?
end p.173

One might approach this question by asking: Would zombies have the key features attributed by thesis *C*? Or at least by asking: Is it conceivable that zombies lack the key features? Note that neither question assumes that zombies are metaphysically possible.

We simply need the assumption that zombies are conceivable, an assumption that type-B materialists typically grant.

One can also approach the question by considering a scenario closer to home. Instead of considering physically identical zombies, we can consider functionally identical zombies: say, functionally identical creatures that have silicon chips where we have neurons and that lack consciousness. Most type-B materialists allow that it is at least an open epistemic possibility that silicon functional isomorphs in the actual world would lack consciousness. We can then ask: Assuming that these functional isomorphs lack consciousness, do they also lack the key features attributed by thesis *C*? If it is conceivable that a functional isomorph lacks these features, then it will almost certainly be conceivable that a physical isomorph lacks these features.

In any case, either physical duplicates that lack the key features are conceivable or they are not. This allows us to set up a master argument against the phenomenal concept strategy, in the form of a dilemma:

1. If $P \& \sim C$ is conceivable, then *C* is not physically explicable.
2. If $P \& \sim C$ is not conceivable, then *C* cannot explain our epistemic situation.
3. Either *C* is not physically explicable, or *C* cannot explain our epistemic situation.

The argument is valid. It has the form of a dilemma, with each premise representing one of the horns. In what follows I will discuss each horn in turn, arguing for the corresponding premise.

First Horn: $P \& \sim C$ Is Conceivable

Premise (1) says that if $P \& \sim C$ is conceivable, then *C* is not physically explicable. The argument for this premise is straightforward. It parallels the original reasoning from the claim that $P \& \sim Q$ is conceivable to the claim that *Q* is not physically explicable. If one can conceive of physical duplicates that lack the key features attributed by thesis *C*, then there will be an explanatory gap between *P* and *C*. That is, there will be no wholly physical explanation that makes transparent why thesis *C* is true. To explain why, in the actual world, creatures with the relevant physical structure satisfy thesis *C*, we will need additional explanatory materials, just as we need such principles to explain why actual creatures with this physical structure are conscious.

Here, again, we are assuming nothing about the relationship between conceivability and possibility. It may be that creatures satisfying $P \& \sim C$ are metaphysically impossible. We are simply assuming a connection between conceivability and explanation. More precisely, we are assuming a connection between conceivability and a certain sort of reductive explanation, the sort that is relevant here: explanation that makes transparent why some high-level truth obtains, given that certain low-level truths obtain. If it is conceivable that the low-level truths obtain without the high-level truths obtaining, then this sort of transparent explanation will fail. The original explanatory gap between consciousness and the physical turns on the

end p.174

absence of just this sort of transparent explanation. If it is conceivable that *P* obtains without *C* obtaining, then we will have just the same sort of explanatory gap between physical processes and the relevant features of phenomenal concepts.

Type-B materialists typically accept this connection between conceivability and transparent explanation, even though they reject the connection between conceivability and possibility. So for now, I will take the connection between conceivability and explanation for granted. Later I will argue that even rejecting the connection will not remove the dilemma for the type-B materialist.

One might think that a proponent of the phenomenal concept strategy *must* take this first horn of the dilemma, as thesis *C* will be a thesis about *phenomenal* concepts. If thesis *C* explicitly requires the existence of phenomenal concepts, and if phenomenal concepts require the existence of phenomenal states, then it is out of the question that zombies could have the features attributed by thesis *C*. If *C* builds in the truth of *Q*, and *P* & ~*Q* is conceivable, then *P* & ~*C* will automatically be conceivable. A physical explanation of the truth of thesis *C* would then be ruled out.

We can avoid this problem by stipulating that thesis *C* should be cast in *topic-neutral* terms: terms that do not explicitly attribute phenomenal states or concepts that refer to them. The restriction to topic-neutral terms allows that thesis *C* may include psychological or epistemological vocabulary, in addition to physical and functional vocabulary. But phenomenal vocabulary is barred. For example, instead of casting thesis *C* as a thesis explicitly about phenomenal concepts, one can cast it as a thesis about *quasi-phenomenal* concepts, where these can be understood as concepts deployed in certain circumstances that are associated with certain sorts of perceptual and introspective processes, and so on. Phenomenal concepts will be quasi-phenomenal concepts, but now it is not out of the question that zombies might have quasi-phenomenal concepts too.

Formulated this way, thesis *C* will then say that quasi-phenomenal concepts have certain properties, such as being recognitional concepts without contingent modes of presentation. We can likewise appeal to quasi-phenomenal concepts in characterizing our epistemic situation with regard to consciousness. This allows the possibility that even if consciousness cannot be physically explained, we might be able to physically explain the key psychological features and our epistemic situation. If we could physically explain why we are in such an epistemic situation, we would have done the crucial work in physically explaining the existence of an explanatory gap.

Henceforth, I will take it for granted that thesis *C* should be cast in topic-neutral terms.

The same goes for the characterization of our epistemic situation. Understood this way, it is by no means out of the question that zombies would have quasi-phenomenal concepts with the properties in question, and that *P* & ~*C* is not conceivable, leading to the second horn of the dilemma. That question is no longer prejudiced by building in theses about phenomenology. Rather, the question will turn on the character of the psychological features themselves.¹

end p.175

Of course, it remains possible that even when thesis C is understood in topic-neutral terms, the character of the psychological features involved in C is such that $P \& \sim C$ is conceivable. If so, then the first horn of the dilemma is raised as strongly as ever. On this horn, the relevant psychological features will raise just as much of an explanatory gap as consciousness itself, and an appeal to these features can do little to deflate the explanatory gap.

Second Horn: $P \sim C$ Is Not Conceivable

Premise 2 says that if $P \sim C$ is not conceivable, then C cannot explain our epistemic situation. The case for this premise is not quite as straightforward as the case for premise 1. One can put the case informally as follows:

4. If $P \sim C$ is not conceivable, then zombies satisfy C .
5. Zombies do not share our epistemic situation.
6. If zombies satisfy C but do not share our epistemic situation, then C cannot explain our epistemic situation.
7. If $P \sim C$ is not conceivable, then C cannot explain our epistemic situation.

Strictly speaking, the references to zombies should be put within the scope of a conceivability operator. One can formalize the argument in this fashion, but for now I will use the informal version for ease of discussion.²

Here, premise (6) is simply another application of the connection between conceivability and explanation. Premise (4) might be derived from a principle of completeness about the conceivable (if R is conceivable, then for arbitrary S , either $R \& S$ is conceivable, or $R \sim S$ is conceivable). But in this context, one can also defend (4) more straightforwardly by noting that if the truth of C is transparently explained by P , as the first horn requires, then if we specify that P holds in a conceivable situation, it will follow transparently that C holds in that situation.

The real work in this argument is done by premise (5). This premise amounts to the claim that $P \sim E$ is conceivable, where E characterizes our epistemic situation. To clarify this premise further, one needs to clarify the notion of our epistemic situation.

I will take it that the epistemic situation of an individual includes the truth-values of their beliefs and the epistemic status of their beliefs (as justified or unjustified, and as cognitively significant or insignificant). As before, an epistemic situation (and a sentence E characterizing it) should be understood in topic-neutral terms, so that it does not build in claims about the presence of phenomenal states or

end p.176

phenomenal concepts. We can say that two individuals share their epistemic situation when they have corresponding beliefs, all of which have corresponding truth-value and epistemic status. A zombie will share the epistemic situation of a conscious being if the zombie and the conscious being have corresponding beliefs, all of which have corresponding truth-values and epistemic status. Here, I assume an intuitive notion of correspondence between the beliefs of a conscious being and the beliefs (if any) of its zombie twin. For example, corresponding utterances by a conscious being and its zombie twin will express corresponding beliefs. It is important to note that this notion of correspondence does not require that corresponding beliefs have the same *content*. It is plausible that a nonconscious being such as a zombie cannot have beliefs with exactly the same content as our beliefs about consciousness. But we can nevertheless talk of the zombie's *corresponding* beliefs. So the claim that a zombie and a conscious being share their epistemic situation does not require that their beliefs have the same content. This mirrors the general requirement that epistemic situations be understood in topic-neutral terms.

I will assume here, at least for the sake of argument, that zombies can have beliefs (that is, that it is conceivable that zombies have beliefs). This is by no means obvious. But if zombies cannot have beliefs, then the phenomenal concept strategy cannot get off the ground. If zombies cannot have beliefs, then presumably they cannot possess concepts either, so there will be an explanatory gap between physical processes and the possession of concepts. If so, then there will be an explanatory gap between physical processes and the key features of phenomenal concepts, leading to the first horn of the dilemma. And even if zombies can have concepts with the key features, then as long as they cannot have beliefs, the key features cannot explain our epistemic situation, leading to the second horn of the dilemma. So the assumption that zombies can have beliefs should be seen as a concession to the type-B materialist for the sake of argument.

For a given conscious being with a given epistemic situation as understood above, *E* will be a sentence asserting the existence of a being with that epistemic situation. This sentence will be made true by that being in its original epistemic situation, and it will be made true by any being that shares this epistemic situation in the sense specified above. Premise 5, the claim that zombies do not share our epistemic situation, can be understood as the claim that $P \neg E$ is conceivable, where *E* characterizes the epistemic situation of an actual conscious being. That is, it is the claim that (it is conceivable that) zombies' beliefs differ in their truth-value or their epistemic status from the corresponding beliefs of their actual conscious twins.

Why think that zombies do not share our epistemic situation? The first reason for this is intuitive. On the face of it, zombies have a much less accurate self-conception than conscious beings do. I believe that I am conscious, that I have states with remarkable qualitative character available to introspection, that these states resist transparent reductive explanation, and so on. My zombie twin has corresponding beliefs. It is not straightforward to determine just what content these beliefs might possess. But there is a strong intuition that these beliefs are false, or at least that they are less justified than my beliefs.

One can develop this intuitive consideration by considering a zombie's utterances of sentences such as "I am phenomenally conscious." It is not clear exactly
end p.177

what a zombie asserts in asserting this sentence. But it is plausible that the zombie does not assert a truth.

Balog (1999) suggests that the zombie does assert a truth, as its term “phenomenal consciousness” will refer to a brain state. This seems to give implausible results, however. We can imagine a debate in a zombie world between a zombie eliminativist and a zombie realist:

Zombie Eliminativist: “There's no such thing as phenomenal consciousness.”

Zombie Realist: “Yes, there is.”

Zombie Eliminativist: “We are conscious insofar as ‘consciousness’ is a functional concept, but we are not conscious in any further sense.”

Zombie Realist: “No, we are conscious in a sense that is not functionally analyzable.”

When such a debate is held in the actual world, the type-B materialist and the property dualist agree that the zombie realist is right, and the zombie eliminativist is wrong. But it is plausible that in a zombie scenario, the zombie realist would be wrong, and the zombie eliminativist would be right. If so, then where we have true beliefs about consciousness, some corresponding beliefs of our zombie twins are false, so that zombies do not share our epistemic situation.

Still, because judgments about the truth-value of a zombie's judgments are disputed, we can also appeal to a different strategy, one that focuses on the nature of our knowledge compared to a zombie's knowledge. Let us focus on the epistemic situation of Mary, upon seeing red for the first time. Here, Mary gains cognitively significant knowledge of what it is like to see red, knowledge that could not be inferred from physical knowledge. What about Mary's zombie twin, Zombie Mary? What sort of knowledge does Zombie Mary gain when she emerges from the black-and-white room?

It is plausible that Zombie Mary at least gains certain abilities. For example, upon seeing a red thing, she will gain the ability to perceptually classify red things together. It is also reasonable to suppose that Zombie Mary will gain certain *indexical* knowledge, of the form *I am in this state now*, where *this state* functions indexically to pick out whatever state she is in. But this knowledge is analogous to trivial indexical knowledge of the form *It is this time now*, and is equally cognitively insignificant. There is no reason to believe that Zombie Mary will gain cognitively significant introspective knowledge, analogous to the cognitively significant knowledge that Mary gains. On the face of it, there is nothing for Zombie Mary to gain knowledge of. For Zombie Mary, all is dark inside, so even confronting her with a new sort of stimulus will not bring about new significant introspective knowledge.

If this is right, then Zombie Mary does not share Mary's epistemic situation. In addition to Mary's abilities and her indexical beliefs, Mary has significant knowledge of what it is like to see red, knowledge not inferable from her physical knowledge. But Zombie Mary does not have significant non-indexical knowledge that corresponds to Mary's knowledge. If so, then Zombie Mary does not share Mary's epistemic situation.

One can also bring out the contrast by considering a case somewhat closer to home.

Balog (1999) appeals to hypothetical conscious humans called “Yogis,” who have the ability to refer directly to their brain states by deploying direct recognitional

concepts of those states, even when those states have no associated phenomenal quality. She suggests that zombies likewise might have direct recognitional knowledge of their brain states by deploying a recognitional concept analogous to a Yogi's.

Even if Yogi concepts like this are possible, however, it is clear that they are nothing like phenomenal concepts. A Yogi going into a new brain state for the first time *might* sometimes acquire a new recognitional concept associated with that state. But a Yogi will not acquire new cognitively significant knowledge that is analogous to Mary's phenomenal knowledge. At best, a Yogi will acquire trivial knowledge, which we might express roughly as "that sort of brain state is that sort of brain state." So even if Zombie Mary can have a recognitional concept like this, she will still not have an epistemic situation like Mary's.

(I think a Yogi's concept is probably best understood as a response-dependent concept: if the concept is *flurg*, it is a priori for the Yogi that a flurg is whatever brain state normally triggers flurg-judgments. Once a Yogi discovers that brain state *B* triggers these judgments, he will know that a flurg is an instance of *B*, and there will be no further question about flurges. This contrasts with a phenomenal concept: once we discover that our phenomenal redness judgments are typically triggered by brain state *B*, we will still regard the question of the nature of phenomenal redness as wide open. This difference between response-dependent concepts and phenomenal concepts tends to further undercut Balog's suggestion that Yogi's concepts are just like phenomenal concepts.)

If the above is correct, then $P \sim E$ is conceivable, and premise 5 is correct. When this is combined with premises 4 and 6, the conclusion follows. That is: if $P \sim C$ is not conceivable, then Zombie Mary has the psychological features attributed by *C*, but she does not share Mary's epistemic situation. So the psychological features attributed by *C* cannot explain Mary's epistemic situation, and more generally, cannot explain our epistemic situation with respect to consciousness.

Summary

We can summarize the arguments above more briefly as follows.

1. $P \sim E$ is conceivable
2. If $P \sim E$ is conceivable, then $P \sim C$ is conceivable or $C \sim E$ is conceivable.
3. If $P \sim C$ is conceivable, *P* cannot explain *C*.
4. If $C \sim E$ is conceivable, *C* cannot explain *E*.
5. P cannot explain *C* or *C* cannot explain *E*.

Premise (1) is supported by the considerations about Zombie Mary above. Premise (2) is a plausible consequence of the logic of conceivability. Premises (3) and (4) are applications of the connection between conceivability and explanation. The conclusion says that *C* cannot satisfy the constraints laid out in the general requirements for the phenomenal concept strategy. The argument here is general, applying to any candidate for *C*. It follows that the phenomenal concept strategy cannot succeed: no psychological features are simultaneously physically explicable and able to explain the distinctive epistemic gaps in the phenomenal domain.
end p.179

Reactions

Proponents of the phenomenal concept strategy may react to this argument in one of four ways. First, they may accept that *P* cannot explain *C* but hold that the phenomenal concept strategy still has force. Second, they may accept that *C* cannot explain *E* (at least as I have construed *E*) but hold that the phenomenal concept strategy still has force. Third, they may deny that *P*→*E* is conceivable and hold that Zombie Mary shares the same epistemic situation as Mary. Fourth, they may deny the connection between conceivability and explanation. (Each of these reactions has been suggested in discussions I have had with type-B materialists, with the first and third reactions being more common than the second and fourth.) In what follows, I will discuss each of the reactions in turn.

Option 1: Accept That *P* Cannot Explain *C*

The first response adopts what we might call the “thick phenomenal concept” strategy. On this approach, proponents appeal to features of phenomenal concepts that are thick enough to explain our distinctive epistemic situation with respect to consciousness but are too thick to be physically explained.

An example of such an approach may be the proposal that phenomenal concepts involve a direct acquaintance with their referent of a sort that discloses an aspect of their referent's intrinsic nature. Such a proposal may well help to explain the distinctive epistemic progress that Mary makes and that Zombie Mary does not make: Mary has concepts that involve direct acquaintance with their referents, whereas Zombie Mary does not. But the very fact that Mary has such concepts and that Zombie Mary does not suggests that this feature of phenomenal concepts cannot be physically explained. The proposal requires a special psychological feature (acquaintance) whose existence one would not predict from just the physical/functional structure of the brain.

The obvious problem here is the problem mentioned before. On this account, even if there is a sort of explanation of the explanatory gap in terms of features of phenomenal concepts, the explanatory gap recurs just as strongly in the explanation of phenomenal concepts themselves. Because of this, the strategy may make some progress in *diagnosing* the explanatory gap, but it will do little to deflate the gap.

end p.180

A proponent may suggest that just as the first-order explanatory gap can be explained in terms of second-order features of phenomenal concepts, the second-order explanatory gap concerning phenomenal concepts can be explained in terms of third-order features of our concepts of phenomenal concepts, and so on. Alternatively, an opponent may suggest that the second-order explanatory gap can be explained in terms of the same second-order features of phenomenal concepts that explain the first-order explanatory gap. The first move here obviously leads to a regress of explanation, and the second move leads to a circular explanation. Explanatory structures of this sort can be informative, but again they will do nothing to deflate the explanatory gap unless the chain of explanation is at some point grounded in physical explanation.

A proponent may also suggest that to require that the key psychological features be physically explicable is to set the bar too high. On this view, all that is needed is a psychological explanation of the epistemic gap that is *compatible* with the truth of physicalism, not one that is itself transparently explainable in physical terms. However, an opponent will now question the compatibility of the account with the truth of physicalism. Just as the original explanatory gap gave reason to think that consciousness is not wholly physical, the new explanatory gap gives reason to think that phenomenal concepts are not wholly physical.

At this point, the proponent may respond by saying that ontological physicalism is compatible with the existence of explanatory gaps. But now we are back where we started, before the phenomenal concept strategy came in. Antiphysicalists argue from an epistemic gap to an ontological gap. The phenomenal concept strategy as outlined earlier was supposed to ground the rejection of this inference by showing how such epistemic gaps can arise in a purely physical system. If successful, the strategy would help to *justify* the claim that the epistemic gap is compatible with ontological physicalism, and so would lend significant support to type-B materialism. But the weaker version of the strategy outlined above can give no such support. On this version, the proponent needs *independent* grounds to reject the inference from an explanatory gap to an ontological gap. If the proponent has no such grounds, then the phenomenal concept strategy does nothing to provide them. An opponent will simply say that the explanatory gap between physical processes and phenomenal concepts provides all the more reason to reject physicalism. If the proponent already has such grounds, on the other hand, then the phenomenal concept strategy is rendered redundant. Either way, the strategy will play no role in supporting type-B materialism against the antiphysicalist.

This limitation does not entail that the limited version of the phenomenal concept strategy is without interest. Even if it does not *support* a type-B materialist view, we can see this sort of account of phenomenal concepts as helping to *flesh out* a type-B materialist view by giving an account of what phenomenal concepts might be like under the assumption that type-B materialism is true. If we have independent reasons to be type-B materialists, we may then have reason to suppose that phenomenal concepts work as the account suggested. And if we have some independent method of deflating the original explanatory gap, then presumably this method may also apply to the new explanatory gap. For example, if a type-B materialist accepts an explanatorily primitive identity between certain physical/functional properties and phenomenal properties, she may also

accept an explanatorily primitive identity between certain physical/functional properties and the properties of phenomenal concepts. But insofar as one has reasons to reject type-B materialism, the phenomenal concept strategy will do nothing to undermine these reasons.

(Note that I am not arguing in this chapter that type-B materialism is false. I have done that elsewhere. Here I am simply arguing that the phenomenal concept strategy provides no support for type-B materialism and provides no grounds for rejecting arguments from the epistemic gap to an ontological gap.)

Overall, I think that accepting an explanatory gap between physical processes and phenomenal concepts is the most reasonable reaction to the arguments above for a type-B materialist. To accept such a gap does not immediately rule out the
end p.181

truth of type-B materialism, and the account of phenomenal concepts may help in elaborating the position. But now the phenomenal concept strategy does nothing to *support* type-B materialism against the antimaterialist. To resist antimaterialist arguments, and to deflate the significance of the explanatory gap, the type-B materialist must look elsewhere.

Option 2: Accept That *C* Does Not Explain *E*, But Hold That It Explains a Reconstructed *E*

The second possible reaction for a type-B materialist is to embrace the second horn of the dilemma, accepting that the key psychological features that they appeal to do not explain our epistemic situation, at least as I have construed that epistemic situation. We might think of this as a “thin phenomenal concepts” strategy. Here, the psychological features in question are tame enough to be physically explained, but they are not powerful enough to explain the full-blown epistemic gaps associated with consciousness.

The problem with this strategy is the same as the problem for the first strategy. Because it leaves a residual explanatory gap, it does little to close the original explanatory gap. The issues that come up here are similar to the issues under the first reaction, so I will not go over them again. If anything, this reaction is less attractive than the first reaction because an account of phenomenal concepts that cannot explain our epistemic situation with regard to consciousness would seem to have very little to recommend it.

There is a version of this reaction that is worth attending to, however. This version concedes that the key psychological features in question cannot explain our full epistemic situation *as I have defined it*, but asserts that the features can explain our epistemic situation in a narrower sense, where it is this sense that is crucial to explaining away the explanatory gap. In particular, a proponent may suggest that I raised the bar unnecessarily high by stipulating that our epistemic situation includes the *truth-values* of our beliefs, and by including their status as *knowledge*. It may be suggested that there is a sense in which truth-value is external to our epistemic situation, and that the phenomenal concept strategy needs only to explain our epistemic situation more narrowly construed.

I can think of three main versions of this strategy. A proponent may suggest: (1) that a physically explicable account of phenomenal concepts can explain the *justification* of our phenomenal beliefs; or (2) that such an account can explain the *inferential disconnection*

between our physical and phenomenal beliefs, including the fact that the latter are not deducible from the former, for example (this suggestion meshes especially well with Hill's account of phenomenal concepts in terms of dual conceptual roles); or (3) that such an account can explain the *existence* of our phenomenal beliefs and of associated beliefs, such as the belief in an explanatory gap. In each of these cases, proponents may claim that corresponding features will be present in zombies, so that there is no obstacle to a physical explanation.

I think that each of these strategies is interesting, but each suffers from the same problem.

To restrict the ambition of the phenomenal concept strategy in this way undercuts its force in supporting type-B materialism. Recall that the strategy is intended to resist the antiphysicist's inference from an epistemic gap to an ontological gap by showing how the relevant epistemic gap may exist even if physicalism is true. In the antiphysicist's arguments, the relevant epistemic gap (from which an ontological gap is inferred) is characterized in such a way that truth and knowledge are essential. For example, it is crucial to the knowledge argument that Mary gains new factual *knowledge* or, at least, new true beliefs. It is crucial to the conceivability argument that one can conceive beings that lack phenomenal states that one actually has. And it is crucial to the explanatory gap that one has cognitively significant knowledge of the states that we cannot explain. If one characterized these gaps in a way that were neutral on the truth of phenomenal beliefs, the arguments would not get off the ground. So truth-value is essential to the relevant epistemic gaps. If so, then to undercut the inference from these gaps to an ontological gap, the phenomenal concept strategy needs to show how the relevant truth-involving epistemic gaps are consistent with physicalism. The strategies above do not do this, so they do nothing to undercut the inference from the epistemic gap to an ontological gap. Perhaps proponents could augment their explanation of the narrow epistemic situation with an additional element that explains why the relevant beliefs are true and qualify as knowledge. For example, one might augment it with an explanation (perhaps via a causal theory of reference?) of why phenomenal beliefs refer to physical states and an explanation (perhaps via a reliabilist theory of knowledge?) of why such beliefs constitute knowledge. However, such an augmented explanation is now subject to the original dilemma. If the account applies equally to a zombie (as might be the case for simple causal and reliabilist theories, for example), then it cannot account for the crucial epistemic differences between conscious beings and zombies. And if it does not apply equally to a zombie (if it relies on a notion of acquaintance, for example), then crucial explanatory elements in the account will not be physically explainable.

So I think that none of these strategies gives any support to type-B materialism. Each of them deserves brief discussion in its own right, however. For example, it is worth noting that strategy (1), involving justification, has a further problem in that it is plausible that Mary's introspective beliefs have a sort of justification that Zombie Mary's corresponding beliefs do not share. One could make this case by appealing to the widely accepted view that conscious experience makes a difference to the justification of our perceptual and introspective beliefs. Or one could make it by considering the scenario directly: whereas Mary's belief that she is currently conscious and having a color experience is plausibly justified with something approaching Cartesian certainty, there is a strong intuition that Zombie Mary's corresponding belief is not justified to the same extent, if it is justified at

all. If so, then a physically explicable account of phenomenal concepts cannot explain even the justificatory status of Mary's phenomenal knowledge.

The second strategy, involving inferential disconnection, does not have this sort of problem, as it is plausible that a zombie's physical and quasi-phenomenal beliefs are no more inferentially connected than a conscious being's beliefs. Here, the main problem is that given above. Whereas the inferential disconnection strategy may physically explain an inferential disconnection between physical and phenomenal
end p.183

beliefs, the antiphysicist's crucial epistemic gap involves a disconnection between physical and phenomenal *knowledge*. This strategy does not help to reconcile this crucial epistemic gap with physicalism, so it lends no support to type-B materialism. At best, it shows that zombie-style analogs of phenomenal beliefs (inferentially disconnected from physical beliefs) are compatible with physicalism, but this is something that we knew already.

The most interesting version of strategy (3) is the one that appeals to phenomenal concepts to explain our belief in an epistemic gap (including our belief that Mary gains new knowledge, that zombies are conceivable, and that there is an explanatory gap). For the reasons given above, this strategy cannot help the type-B materialist undermine the *inference* from an epistemic gap to an ontological gap. However, one might think that it helps undermine the premise of that inference by explaining why the belief in such a gap is to be predicted even if no such gap exists. This is an important strategy, but it is one more suited to a type-A materialist than to a type-B materialist. The type-B materialist agrees with the antiphysicist, against the type-A materialist, on the datum that there *is* an epistemic gap (e.g., that zombies are conceivable, that Mary gains new phenomenal knowledge, and that there is an explanatory gap). Given this datum, and given that the inference from an epistemic gap to an ontological gap is unchallenged by this strategy, then the strategy does nothing to support type-B materialism against the antiphysicist.

Option 3: Assert That Zombies Share Our Epistemic Situation

The third reaction is to assert that zombies share our epistemic situation. Where we have beliefs about consciousness, zombies have corresponding beliefs with the same truth-values and the same epistemic status. And where Mary acquires new phenomenal knowledge on seeing red for the first time, Zombie Mary acquires new knowledge of a precisely analogous sort. If this is right, then the crucial features of phenomenal concepts might simultaneously be physically explicable and able to explain our epistemic situation. Of course, a zombie's crucial beliefs will not be *phenomenal* beliefs, and Zombie Mary's crucial knowledge will not be *phenomenal* knowledge. Zombies have no phenomenal states, so they cannot have true beliefs that attribute phenomenal states to themselves, and they cannot have first-person phenomenal knowledge. Instead, the proponent of this strategy must conceive of zombies as attributing some *other* sort of state to themselves. We might think of these states as “schmenomenal states,” and the corresponding beliefs as “schmenomenal beliefs.” Schmenomenal states stand to phenomenal states roughly as

“twater,” the superficially identical liquid on Twin Earth, stands to water: schmenomenal states are not phenomenal states, but they play a role in zombies' lives that is analogous to the role that phenomenal states play in ours. In particular, on this proposal, a zombie's schmenomenal beliefs have the same truth-value and epistemic status as a non-zombie's phenomenal beliefs.

One might worry that in a type-B materialist view, schmenomenal states must be the same as phenomenal states, since both are identical to the same underlying physical states. In reply, one can note that the discussion of zombies falls within
end p.184

the scope of a conceivability operator, and the type-B materialist allows that although physical states are identical to phenomenal states, it is at least conceivable that they are not so identical. The zombie scenario will presumably be understood in terms of conceiving that the same physical states are identical to nonphenomenal (schmenomenal) states instead. To avoid this complication, one might also conduct this discussion in terms of a functionally identical silicon zombie, rather than in terms of a physically identical zombie. Then the type-B materialist can simply say that ordinary humans have neural states that are identical to phenomenal states, whereas silicon zombies have silicon states that are identical to schmenomenal states. On the current view, silicon zombies will have schmenomenal knowledge that is epistemically analogous to humans' phenomenal knowledge.

This proposal might be developed in two different ways: either by deflating the phenomenal knowledge of conscious beings or by inflating the corresponding knowledge of zombies. That is, a proponent may argue either that Mary gains *less* new knowledge than I suggested earlier or that Zombie Mary gains *more* new knowledge than I suggested earlier. Earlier, I argued that Mary gains new cognitively significant non-indexical knowledge, whereas Zombie Mary does not. The deflationary strategy proposes that Mary gains no such knowledge; the inflationary strategy proposes that Zombie Mary gains such knowledge, too.

The deflationary strategy will presumably involve the claim that the only new factual knowledge that Mary gains upon seeing red for the first time is *indexical* knowledge. That is, Mary gains knowledge of the form “I am in *this* state now,” where “*this* state” picks out the state that she happens to be in: presumably some sort of neural state. According to this proposal, Zombie Mary gains analogous knowledge, also of the form “I am in *this* state now,” where “*this* state” picks out the state she happens to be in: presumably a neural state or a silicon state. There seems to be no problem in principle with the idea that Zombie Mary could gain indexical knowledge of this sort, at least if a zombie can have knowledge at all. This strategy meshes particularly well with the proposal that phenomenal concepts are a species of indexical concept.

In response, I think there is good reason to accept that Mary gains more than indexical knowledge. I have made this case elsewhere (Chalmers 2003a), so I will just recapitulate it briefly here. First, there is a sense in which indexical knowledge is perspective-dependent, and vanishes from an objective perspective. For me, full objective knowledge is incomplete unless I know that I am David Chalmers, but no one else with full objective knowledge can be ignorant of the fact that I am David Chalmers in this way. The same

goes for indexical knowledge of my current time and location: no one with full objective knowledge can be ignorant of this in the way that I can be ignorant. Mary's indexical knowledge that *this* brain state is such-and-such brain state is of the same sort: that is, no one else with full physical knowledge can be ignorant of this in the way that Mary can be ignorant. But Mary's phenomenal knowledge of what it is like for her to see a red tomato is not like this. Other beings with full physical knowledge can be ignorant of what it is like for Mary to see a tomato, just as Mary was ignorant before she saw the tomato, regardless of their perspective or the brain states they happen to be in. This strongly suggests that Mary's phenomenal knowledge is not indexical knowledge.
end p.185

Second, just as Mary gains nontrivial knowledge that such-and-such is what it is like to see red, where “such-and-such” corresponds to her deployment of a phenomenal concept, she also gains nontrivial indexical knowledge that *this* state is such-and-such, where “*this* state” corresponds to an indexical concept picking out whatever phenomenal state she happens to be in, and “such-and-such” again corresponds to her deployment of a phenomenal concept. This knowledge is cognitively significant knowledge that Mary gains upon introspection. But this knowledge involves the deployment of an indexical concept on one side of an identity, and Mary's crucial phenomenal concept on the other side. Again, this strongly suggests that the phenomenal concept is distinct from the indexical concept, and that Mary's cognitively significant knowledge *I am in such-and-such state now* is distinct from her trivial indexical knowledge *I am in this state now*. If so, then Mary gains more than this indexical knowledge, and the deflationary strategy fails.

The inflationary strategy involves the proposal that just as Mary gains cognitively significant non-indexical knowledge involving phenomenal concepts, Zombie Mary gains analogous cognitively significant non-indexical knowledge involving schmenomenal concepts. So where Mary gains significant knowledge of the form *Tomatoes cause such-and-such phenomenal state, I am in such-and-such phenomenal state, and This state is such-and-such phenomenal state*, Zombie Mary gains significant knowledge of the form *Tomatoes cause such-and-such schmenomenal state, I am in such-and-such schmenomenal state, and This state is such-and-such schmenomenal state*. Zombie Mary's new beliefs have the same truth-value, the same epistemic status, and the same epistemic connections as Mary's corresponding beliefs.

Here, the natural response is that this scenario is simply not what we are conceiving when we conceive of a zombie. *Perhaps* it is possible to conceive of a being with another sort of state—call it “schmonsciousness”—to which it stands in the same sort of epistemic relation that we stand in to consciousness. Schmonsciousness would not be consciousness, but it would be epistemically just as good. It is by no means obvious that a state such as schmonsciousness is conceivable, but it is also not obviously inconceivable. However, when we ordinarily conceive of zombies, we are not conceiving of beings with something analogous to consciousness that is epistemically just as good. Rather, we are conceiving of beings with nothing epistemically analogous to consciousness at all. Put differently: when we conceive of zombies, we are not conceiving of beings whose inner life is as rich as ours, but different in character. We are conceiving of beings whose

inner life is dramatically poorer than our own. And this difference in inner lives makes for dramatic difference in the richness of our introspective knowledge. Where we have substantial knowledge of our phenomenal inner lives, zombies have no analogous introspective knowledge: there is nothing analogous for them to have introspective knowledge of.

Perhaps a zombie can have a sort of introspective knowledge of some of its states: its beliefs and desires, say, or its representations of external stimuli. But this sort of introspective knowledge is not analogous to our phenomenal introspective knowledge. Rather, it is analogous to our nonphenomenal introspective knowledge.

end p.186

knowledge. Phenomenology is not all that is available to introspection, and it is not out of the question that zombies could have the sort of nonphenomenal introspective knowledge that we have. But none of this knowledge will have the character of our introspective knowledge of phenomenal states because there is nothing analogous for zombies to introspect.

At this point a proponent might appeal to certain naturalistic theories of the mind: perhaps a functionalist theory of belief, a causal theory of mental content, and/or a reliabilist theory of knowledge. Zombies have the same functional organization as conscious beings and the same reliable causal connections among their physical states, so a proponent could suggest that these theories entail that zombies will have corresponding beliefs with the same epistemic status as ours. It is not obvious that the theories will make this prediction: this depends on whether they are a priori theories that apply to all conceivable scenarios. If they do not, then they do not undermine the conception of zombies whose epistemic status differs from ours. But in any case, to appeal to these theories in this context is to beg the question. Consideration of the Mary situation and related matters gives us good reason to believe that consciousness is relevant to matters such as mental content and epistemic status. It follows that if consciousness is not itself explainable in physical/functional terms, then any entirely physical/functional theory of content or knowledge will be incomplete. If a theory predicts that a nonconscious zombie would have the same sort of introspective knowledge that we do, then this is reason to reject the theory.

The upshot of all this is that the inflationary strategy does not adequately reflect what we are conceiving when we conceive of a zombie. *Perhaps* it is conceivable that a nonconscious duplicate could have some analogous state, schmonsciousness, of which they have analogous introspective knowledge. But it is also conceivable that a nonconscious duplicate would have no such analogous introspective knowledge. And this latter conceivability claim is all that the argument against the phenomenal concept strategy needs.

Option 4: Reject the Link between Conceivability and Explanation

The fourth possible reaction for proponents of the phenomenal concept strategy is to deny the connection between conceivability and explanation. Such proponents might allow that

$P \sim C$ is conceivable, but hold that nevertheless, P explains C . Or they might allow that $C \sim E$ is conceivable, but hold that nevertheless, C explains E .

Of course, everyone should allow that there are *some* sorts of explanation such that explaining B in terms of A is consistent with the conceivability of B without A . For causal explanation, for example, this is precisely what one expects. The crucial claim is that there is a *sort* of explanation that is tied to conceivability in this way, and that this sort of explanation is relevant to the explanatory gap. This is the sort of micro-macro explanation that I earlier called transparent explanation: explanation that makes transparent why relevant high-level truths obtain, given that low-level truths obtain. If it is conceivable that the low-level truths obtain without the high-level obtaining, the explanation will not be transparent in the relevant way. Instead, one will need to appeal to substantive further principles to bridge the divide between the low-level and high-level domain. It is just this sort of transparent explanation that is absent in the original explanatory gap.

An opponent may deny that this sort of transparent explanation is required for a good reductive explanation or that it is present in typical reductive explanations. Or he may at least deny this for a notion of transparent explanation that is strongly tied to conceivability. For example, Block and Stalnaker (2001), Levine (2001), and Yablo (2002) all argue that typical cases of micro-macro explanation—the explanation of water in terms of H_2O , for example—are not associated with an a priori entailment of macro truths by micro truths. If they are right about this, then insofar as the notion of transparent explanation is tied to a priori entailment, it is not required for ordinary micro-macro explanation. I have argued elsewhere (Chalmers and Jackson 2001) that they are not right about this: even in cases such as the relation between microphysics and water, there is a sort of associated a priori entailment, and this sort of entailment is crucial for a good reductive explanation.

It is also worth noting that even if these theorists are right, this will at best undermine a link between one sort of conceivability and explanation. As before, let us say that S is *negatively* conceivable when the truth of S cannot be ruled out a priori. Then the claim that A entails B a priori is equivalent to the claim that $A \sim B$ is negatively conceivable. If these theorists are right, then even “zombie- H_2O ” (Levine's (2001) term for a microphysically identical substance that is not water) will be negatively conceivable, so that ordinary micro-macro explanation of B by A cannot require that $A \sim B$ is negatively conceivable. However, Levine himself notes that there is a different sort of “thick” conceivability such that zombies are conceivable in this sense and zombie- H_2O is not, and he notes that this sort of conceivability is tied to explanation: $A \sim B$ is thickly conceivable if and only if there is an explanatory gap between A and B . If so, we can use this sort of thick conceivability in the previous arguments.

I think that Levine's thick conceivability corresponds closely to what I earlier called *positive* conceivability, which requires a clear and distinct positive conception of a situation that one is imagining. Positive conceivability is arguably the central philosophical notion of conceivability. And it is highly plausible that in cases of ordinary

reductive explanations of B by A , $A \dashv B$ is not positively conceivable: we can form a positive conception of a zombie in a way that we cannot form a positive conception of zombie- H_2O . Furthermore, this positive conceivability seems to be particularly strongly associated with the sense of apparent contingency that goes along with the explanatory gap. So it remains plausible that for the sort of explanation that is relevant here, positive conceivability of $A \dashv B$ entails an explanatory gap between A and B .

An opponent may insist more strongly that no sort of conceivability is tied in this way to micro-macro explanation. She may hold that this sort of explanation simply requires a relevant correlation or a relevant identity between the low-level and high-level domains, whose existence does not require any strong conceptual connection between low-level truths and high-level truths. I think that this gets the character of micro-macro explanation wrong, by failing to account for the sense of transparency in a good micro-macro explanation. But in any case, an opponent of this sort is unlikely to be too worried by the explanatory gap in the first place. If this sort of move works to dissolve the explanatory gap between physical processes and phenomenal concepts, say, then it will work equally well to dissolve the original

end p.188

explanatory gap between physical processes and consciousness. If so, then once again the phenomenal concept strategy is rendered redundant in explaining the explanatory gap. Of course, such a theorist may still appeal to the phenomenal concept strategy to explain the remaining epistemic gaps (such as the conceivability of zombies) in their own right, independent of any connection to the explanatory gap. Here the general idea will be as before: there is no valid inference from these epistemic gaps to an ontological gap because the existence of these epistemic gaps is compatible with physicalism. But as before, an opponent will question the strategy on the grounds that there is as much of an epistemic gap between physical processes and phenomenal concepts (as characterized by the proponent's account), or between phenomenal concepts and our epistemic situation, as there was between physical processes and consciousness. To respond, the opponent must either deny this epistemic gap (which will raise all the previous issues) or give independent reasons to think that the epistemic gap is compatible with physicalism (which will render the phenomenal concept strategy redundant). Either way, the theoretical landscape will be much as before.

(Strictly speaking, there may be one version of this strategy on which the theoretical landscape will differ. A proponent might appeal to phenomenal concepts solely to explain Mary's new knowledge, without using it to explain either the conceivability of zombies or the explanatory gap. If so, then the conceivability of zombies who do not satisfy this account of phenomenal concepts will not raise the usual regress worry. To avoid a residual epistemic gap with the same character as the original epistemic gap, the proponent would simply need to make the case that Mary could know all about the relevant structural features of phenomenal concepts from inside her black-and-white room. Of course, this proponent will then need some other means to deal with the conceivability argument, and with the explanatory gaps posed both by consciousness and by the account of phenomenal concepts.)

In any case, I think that many of the central points of this chapter can also be made directly in terms of explanation, without proceeding first through conceivability. The analysis in terms of conceivability is useful in providing a tool for fine-grained analyses and arguments, and to get a sense of the options in the theoretical landscape. But with these options laid out, one can also make the case directly that any given account of phenomenal concepts will generate either an explanatory gap between physical processes and phenomenal concepts, or between phenomenal concepts and our epistemic situation. I will make this sort of case in the next section.

Applications

I will now look at some specific accounts of phenomenal concepts in light of the preceding discussion. If what has gone before is correct, then any fully specific account of phenomenal concepts will fall into one of two classes. It will be either a “thick” account, in which the relevant features of phenomenal concepts are not physically explainable (although they may explain our epistemic situation), or a
end p.189

“thin” account, in which the relevant features of phenomenal concepts do not explain our epistemic situation (although they may be physically explainable).

I have already discussed the indexical account of phenomenal concepts (of Ismael, O’Dea, Perry, and others) under the third reaction above. For the reasons given there, I think that this account is clearly a thin account; for example, it does not adequately explain the character of Mary’s new cognitively significant knowledge. So there is reason to believe that phenomenal concepts are not indexical concepts.

I have also discussed the dual-conceptual-role account (of Nagel, Hill, McLaughlin, and others), under the second reaction above. If this account is understood in wholly functional terms, involving the distinctness in functional role of certain representations in the brain, then it is clearly a thin account. For reasons discussed earlier, this account may help to explain an inferential disconnection between physical and phenomenal beliefs, but it cannot explain the character of phenomenal *knowledge*. Perhaps the account could be supplemented by some further element to explain this character (for example, postulating a special faculty of sympathetic knowledge), but then the original dilemma will arise once again for the new account.

The “quotational” account (of Block, Chalmers, Papineau, and others) might be understood either as a thin or a thick account, depending on how it is specified. One may understand this either in a “bottom-up” way, in which we start with purely physical/functional materials and make no assumptions about consciousness, or in a “top-down” way, in which we build consciousness into the account from the start. I will examine each of these versions in turn.

The bottom-up version of the quotational account is specified in purely physical/functional terms, without building any assumptions about consciousness. The basic idea will be that there are some neural states *N* (those that correspond to phenomenal states, though we will not assume that) that can come to be embedded in

more complex neural representations by a sort of “quotation” process, which allows the original state to be incorporated as a constituent. Perhaps this will go along with some sort of demonstrative reference to the original neural state, so that the complex state has the form “That state: *N*.” Of course, it is not obvious that one can explain any sort of demonstrative reference in physical/functional terms, but I will leave that point aside. At this point, we can think of the account as an engineer might. If we designed a system to meet the specifications, what sort of results would we expect? In particular, what sort of knowledge of state *N* would one expect? I think the answer is reasonably clear. One would expect a sort of indexical knowledge of the state, of the form “I am in this state now.” But one would not expect any sort of cognitively significant knowledge of the state's intrinsic character. To see this, note that one might design an identical system where state *N* is replaced by a different state *M* (perhaps another neural state, or a silicon state), with different intrinsic properties. From a bottom-up perspective, we would not expect this change to affect the epistemic situation of the subject in the slightest. States *N* and *M* may make a difference to the subject's knowledge by virtue of their functional role, but from an engineering perspective there is no reason to think that the subject has access to their intrinsic character.

end p.190

So the bottom-up version of the quotational account is best understood as a thin account of phenomenal concepts. It may ground a sort of indexical or demonstrative knowledge of neural states, but it cannot ground the sort of significant non-indexical knowledge of internal states that Mary gains on leaving her black-and-white room. In this respect, the bottom-up version of the quotational account seems to be no better off than the indexical account.

On the top-down version of the quotational account, we build consciousness into the account from the start. In particular, we assume that our initial state *Q* is a *phenomenal* state. (It does not matter to what follows whether we assume in addition that *Q* is or is not a neural state, or whether we stay silent on the matter.) We then stipulate a sort of concept-forming process that incorporates phenomenal states as constituents. Perhaps this process will involve a sort of demonstrative reference to the original phenomenal state, so the resulting concept has the form “That state: *Q*.” What sort of results will we then expect?

We *might* then reasonably expect the subject to have some sort of cognitively significant knowledge of the character of *Q*. In general, when we make demonstrative reference to phenomenal states, we can have cognitively significant knowledge of their character. We could also imagine a functionally identical subject who, in place of *Q*, has a different phenomenal state *R*. In this case, one might expect the substitution to affect the subject's epistemic situation: the new subject will have cognitively significant non-indexical knowledge that it is in phenomenal state *R*, which is quite different from the first subject's knowledge.

This top-down version of the quotational account is quite clearly a thick account of phenomenal concepts. By building phenomenal states into the account, it has the capacity to help explain features of our epistemic situation that the bottom-up account cannot. But precisely because the account builds in phenomenal states from the start, it cannot be

transparently explained in physical terms. This version of the account presupposes the special epistemic features of phenomenal states rather than explaining them.

(Papineau's version of the quotational account appears to be a thin version. Papineau discusses a silicon zombie [2002: 125–27] and suggests that it will have semantic and epistemic features analogous to those of a conscious being. His account seems to point in a direction in which the relevant phenomenal knowledge is all a kind of indexical or demonstrative knowledge, although he does not explicitly make this claim or address the objections to it. By contrast, my own version of this sort of account (Chalmers 2003a) is certainly intended as a thick account.)

The recognitional-concept account (of Loar, Carruthers, Tye, and others) can be handled in a similar way. If we understand the concept in a bottom-up way, involving recognitional processes triggered by neural states, what sort of knowledge will we expect? Here, I think we would once again expect a sort of indexical or demonstrative knowledge of the neural states in question, without any cognitively significant knowledge of their intrinsic character. Once again, we would expect that substituting one neural state for another would make no significant difference to a subject's epistemic situation. So this version of the account can be understood as a thin account.

On the other hand, if we understand the account in a top-down way, as involving recognitional concepts triggered by phenomenal states, then one might well expect
end p.191

it to lead to significant knowledge of the character of these states. It is plausible that merely having a phenomenal state enables us to have a conception of its character, by which we can recognize it (or at least, recognize states reasonably similar to it) when it reoccurs, and such that substituting a different phenomenal state will make a difference to our epistemic situation. This top-down account might well capture something about the difference between a conscious being's epistemic situation and a zombie's situation. But again, this account presupposes the existence of consciousness, along with some of its special epistemic features, so the account is clearly a thick account.

(Loar's own account appears to be a thick account. His discussion of phenomenal concepts presupposes both the existence of consciousness and some of its special epistemic features. In particular, his account crucially relies on the thesis that phenomenal states are presented to us under noncontingent modes of presentation, thus enabling significant knowledge of their character. He defends this assumption by saying that the nonphysicalist accepts the thesis, so the physicalist is entitled to it as well. But of course, the thesis poses a special explanatory burden on the physicalist. How can a neural state of a physical system be presented to a subject under a noncontingent mode of presentation, thus enabling significant knowledge of its character? Loar does not say.)

What about Sturgeon's account of phenomenal concepts, according to which phenomenal states constitute their own canonical evidence? I think that this is probably best understood as a thick account. From a bottom-up perspective, would we expect neural states to constitute their own canonical evidence? When zombies deploy their analogs of phenomenal concepts, do they have analogous states that constitute their own canonical evidence? The answer is not entirely obvious, but on the face of it, the more plausible

answer is no. If so, then Sturgeon's account can be seen as a thick account, one that rests on a special epistemic feature of phenomenal concepts.

What do the thick accounts of phenomenal concepts have in common? All of them implicitly or explicitly build in special epistemic features of phenomenal concepts: the idea that phenomenal states present themselves to subjects in especially direct ways, or the idea that simply having a phenomenal state enables a certain sort of knowledge of the state, or the idea that the state itself constitutes evidence for the state. If we build in such features, then we may be able to explain many aspects of our distinctive epistemic situation with respect to consciousness. But the cost is that such features themselves pose an explanatory problem. If these features are powerful enough to distinguish our epistemic situation from that of a zombie, then they will themselves pose as much of an explanatory gap as does consciousness itself.

If one rejects physicalism, there is no obvious problem in accommodating these epistemic features of consciousness. Dualists sometimes postulate an epistemic relation of acquaintance that holds between subjects and their phenomenal states, and that affords knowledge of these states. If necessary, a dualist can simply take this relation as primitive: the dualist is already committed to positing primitive mental features, and this relation may reasonably be taken to be part of the primitive structure of consciousness.

However, this move is not available to a
end p.192

physicalist. The physicalist must either explain the features or accept a further explanatory gap.

Our examination of specific accounts of phenomenal concept reaches a conclusion very much compatible with that of Levine (chap. 8, this volume). It appears that such accounts either build in strong epistemic relations such as acquaintance, which themselves pose problems for physical explanation, or they build in weak epistemic relations such as indexical or demonstrative reference, in which case they cannot explain our epistemic situation with regard to consciousness. The arguments earlier in the chapter suggest that this is not a mere accident of these specific accounts that a better account may evade. Any account of phenomenal concepts can be expected to have one problem or the other. For this reason, the phenomenal concept strategy cannot reconcile ontological physicalism with the explanatory gap.

Acknowledgments

This chapter was greatly influenced by a round-table discussion at the 2002 NEH Summer Institute on Consciousness and Intentionality in which Kati Balog, Ned Block, John Hawthorne, Joe Levine, and Scott Sturgeon, and others took part. I was also influenced by Levine's "Phenomenal Concepts and the Materialist Constraint" (chap. 8, this volume). I first formally presented this chapter at a session on David Papineau's book *Thinking about Consciousness* at the 2003 Pacific APA meeting, and since then have presented it at numerous universities and at workshops in Buenos Aires and Copenhagen. Thanks to all those present on those occasions for very useful reactions.

References

- Aydede, M., and Güzeldere, G. (2005). Cognitive Architecture, Concepts, and Introspection: An Information-Theoretic Solution to the Problem of Phenomenal Consciousness. *Nous* 39: 197–255. [Link ▶](#)
- Balog, K. (1999). Conceivability, Possibility, and the Mind-Body Problem. *Philosophical Review* 108: 497–528. [Link ▶](#)
- Block, N., and Stalnaker, R. (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *Philosophical Review* 108: 1–46. [Link ▶](#)
- Carruthers, P. (2004). Phenomenal Concepts and Higher-Order Experiences. *Philosophy and Phenomenological Research* 68: 316–36.
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Chalmers, D. J. (1999). Materialism and the Metaphysics of Modality. *Philosophy and Phenomenological Research* 59: 473–96. [Link ▶](#)
- Chalmers, D. J. (2002). Does Conceivability Entail Possibility? In *Conceivability and Possibility*, ed. T. Gendler and J. Hawthorne. Oxford: Oxford University Press: 145–200.
- Chalmers, D. J. (2003a). The Content and Epistemology of Phenomenal Belief. In *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic: 220–72. Oxford: Oxford University Press.
- Chalmers, D. J. (2003b). Consciousness and Its Place in Nature. In *The Blackwell Guide to the Philosophy of Mind*, ed. P. Stich and T. Warfield. Oxford: Blackwell. Reprinted in *The Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 247–72. New York: Oxford University Press, 2002.
- end p.193
- Chalmers, D. J., and Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110: 315–61.
- Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown.
- Hill, C. S. (1997). Imaginability, Conceivability, Possibility, and the Mind-Body Problem. *Philosophical Studies* 87: 61–85. [Link ▶](#)
- Hill, C. S., and McLaughlin, B. P. (1999). There Are Fewer Things in Reality Than Are Dreamt of in Chalmers' Philosophy. *Philosophy and Phenomenological Research* 59: 445–54. [Link ▶](#)
- Ismael, J. (1999). Science and the Phenomenal. *Philosophy of Science* 66: 351–69. [Link ▶](#)
- Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Clarendon Press.
- Jackson, F. (2003). Mind and Illusion. In *Minds and Persons*: Royal Institute of Philosophy Supplement 53, ed. A. O'Hear: 251–71. Cambridge: Cambridge University Press.
- Levine, J. (1983). Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64: 354–61.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press. [Link ▶ OSO X-Reference](#)

Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview. Revised version in *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzeldere: 597–616. Cambridge: MIT Press, 1997.

Nagel, T. (1974). What Is It Like to Be a Bat? *Philosophical Review* 4: 435–50. [Link ▶](#)

O'Dea, J. (2002). The Indexical Nature of Sensory Concepts. *Philosophical Papers* 31: 169–81.

Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press.

[Link ▶ OSO X-Reference](#)

Perry, J. (1979). The Problem of the Essential Indexical. *Nous* 13: 3–21.

Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press.

Stoljar, D. (2005). Physicalism and Phenomenal Concepts. *Mind and Language* 20: 469–94. [Link ▶](#)

Sturgeon, S. (1994). The Epistemic Basis of Subjectivity. *Journal of Philosophy* 91: 221–35.

Tye, M. (2003). A Theory of Phenomenal Concepts. In *Minds and Persons*, ed. A. O'Hear: 91–105. Cambridge: Cambridge University Press.

White, S. (1986). Curse of the Qualia. *Synthese* 68: 333–68. [Link ▶](#)

Yablo, S. (2002). Coulda, Woulda, Shoulda. In *Conceivability and Possibility*, ed. T. Gendler and J. Hawthorne. Oxford: Oxford University Press: 441–92.
end p.194

ten Direct Reference and Dancing Qualia

John Hawthorne

A direct reference theory for a term holds that the semantic content of that term is the referent itself. One important group of philosophers defends direct reference accounts for ordinary proper names and demonstratives, attempting to disarm standard Fregean complaints that their account generates an unacceptable rift between semantic content and cognitive significance. A second group accepts the standard style of criticism for direct reference theories of ordinary singular terms but still maintains a direct reference theory for a special class of terms whose reference lies within the Cartesian theater of phenomenal experience.¹ This latter group of philosophers is the concern of this chapter. I shall undertake to expose a tension between this second species of direct reference theory and standard antiphysicalist views of phenomenal experience. A single thought experiment will serve as the centerpiece.

Direct Phenomenal Concepts

There are plenty of ways of thinking and talking about qualia. “Those qualia that I had yesterday,” “The feel of tension headaches,” “God's favorite type of quale,” and “What it's like to be a bat” are all perfectly good ways for speaking of the phenomenal world. Yet there appears to be an especially intimate way of forming a conception of phenomenal experience. I can focus on a particular experience token
end p.195

and form a conception of a phenomenal kind that the experience falls under. I might express such a conception as follows: “*Thus* is one of the ways that I am feeling right now.” One should not be misled into assimilating such concepts to bare demonstratives because each involves a *conceptualization* of a phenomenal type and brings with it various capacities to discriminate sameness and difference that would not be forthcoming from any “blind” ostension. Following David Chalmers, let us call this kind of concept a “direct phenomenal concept.”²

It is direct phenomenal concepts to which our second group of philosophers attaches a direct reference theory: the semantic value of a direct phenomenal concept is the presented phenomenal kind itself. The view is self-consciously Russellian in inspiration: recall that Russell believed that although the content of, say, “Bismarck” could not be Bismarck himself, there is a special class of logically proper names (and basic predicates) that stand for the objects of direct acquaintance, which makes them appropriate to directly referential semantic treatment.³

Qualia and the Physical

What is the relation between phenomenal facts and physical facts? The literature divides between those who take phenomenal facts to be necessitated by physical facts (the most straightforward version of that view being one according to which phenomenal kinds are identical to certain physical kinds) and those who take phenomenal facts to be only contingently related to physical facts: the nomic connections that bridge the gulf between the physical and phenomenal worlds could have been different. I shall be primarily concerned with the second, antiphysicalist, position here.

Famously, antiphysicalists are happy to admit the possibility of zombies, beings that duplicate us in all physical respects, but which lack a phenomenal life.⁴ But given the postulated contingency of the physical to phenomenal connection, there are plenty of other kinds of physical duplicates that ought to engage the imagination. There is a physical duplicate of me who shares my qualia on Mondays, Wednesdays, and Fridays but who is a zombie the rest of the time.⁵ There is a physical duplicate of me who alternates between three seconds of being a zombie and three seconds of phenomenal life. There is a physical duplicate of me whose
end p.196

phenomenal life duplicates mine except for the fact that, for one brief interval, the qualia that it has when complaining of pain are of the same type as certain qualia that I have during certain moments of intense pleasure.

Reflecting on such information, the antiphysicist is naturally led to allow for episodes of what Chalmers calls “dancing qualia,” in which there are marked shifts in phenomenal experience in a subject who does not believe that anything unusual has occurred in his experiential life. The periodic zombie will believe that he has always had experience during waking hours. The complaining duplicate with the once-in-a-lifetime substitution will not register that anything strange is going on.

A Case

Fred and Twin Fred are told that the right-hand side of their phenomenal field is going to “dance” during a given period of time. More specifically, Fred and Twin Fred are told that on three or four occasions during a five-minute interval, there will be a sudden change in the type of qualia that occupy the right-hand side of their phenomenal field without their knowing that a change has occurred. Fred is lied to: during the relevant five minutes, no qualia dancing goes on. In fact, his phenomenal theater consists of a continuous expanse of phenomenal red throughout the period. Twin Fred is not lied to. (The reader is free to choose between a version of the story where God or a superscientist told him knowledgeably that his qualia were going to dance or else someone told him on the basis of a very lucky guess.) At various times during the five minutes, the right side of his phenomenal field switches from phenomenal red to phenomenal blue and then back again. Twin Fred, being a physical duplicate of Fred, is altogether unable to say when any such change occurs. After two minutes, each Fred attends to the left side of his phenomenal field, forms a direct phenomenal concept of the phenomenal type that is present there, and utters the triviality, “Thus is Thus,” where the newly minted phenomenal concept is expressed twice over.⁶ After three minutes, Fred and Twin Fred form a direct phenomenal concept of the phenomenal type on the left of their phenomenal field and a direct phenomenal concept of the phenomenal type on the right of their phenomenal field and, albeit hesitantly, make an identity claim, which they express with the same string of phonemes: “Thus is thus.”^{7, 8} Let us suppose that, once again, they both express truths. (The qualia were dancing in the right direction for Twin Fred at that moment.) After four minutes, they again form direct phenomenal concepts of the kinds displayed on the left and right and, once again, express an identity claim with “Thus is Thus.” This time Fred expresses a truth, and Twin Fred does not. After four minutes it is phenomenal blue that occupies the right-hand side of Twin Fred's phenomenal field.

Preliminary Discussion

Consider the standard Fregean case against direct reference theory:⁹ Someone who considers the thought that Hesperus is Hesperus is considering an utter triviality about

which he can be a priori certain. Someone who considers the thought that Hesperus is Phosphorus is entertaining a thought that cannot be verified a priori, one which instead is judged true or false on the basis of empirical inquiry. Hence the thought that Hesperus is Hesperus is a different thought from the thought that Hesperus is Phosphorus. The same point can be couched in the language of epistemic possibility: ¹⁰ it is epistemically possible that $\neg$ (Hesperus is Phosphorus), but it is not epistemically possible that $\neg$ (Hesperus is Hesperus).

It is not my concern here to defend Fregean arguments against direct reference theory. Instead, I wish to press a point that should by now be obvious: if one reckons Fregean arguments to refute a direct reference theory for ordinary proper names, then one should take a dim view of those antiphysicalists who propose a direct reference theory for direct phenomenal concepts. The familiar distinctions that apply to Hesperus/Phosphorus thoughts can be reenacted at the level of those concepts.

Consider: The first judgment made by Fred and Twin Fred has a priori security. Just as it is not coherently conceivable that Hesperus is not Hesperus, Fred and Twin Fred would not find it coherently conceivable that their first judgment is false. ¹¹ Matters are quite different when it comes to Fred and Twin Fred's second and third judgments. They are both understandably hesitant with regard to those judgments. And they *ought* to be hesitant. The propriety of hesitance is made manifest by Twin Fred's error. None of the relevant four judgment tokens express a priori knowable thoughts. In each case, it is epistemically possible that the judgment is false. But if the contents of those thoughts are individuated in line with a direct reference theory

end p.198

for direct phenomenal concepts, then they will turn out to be the very same thoughts as the first thought entertained by Fred and Twin Fred, respectively. The salient distinction between an a priori knowable thought and epistemically risky thoughts will have been obliterated. Those who reckon the original Fregean line of argument persuasive should reckon the preceding reflections equally cogent. ¹²

If he is to maintain his position, the antiphysicalist Russellian will have to insist that although there is no good a priori justification for the claim that Hesperus is Phosphorus, there is always a powerful justification available for any thought that expresses the proposition expressed by Fred's second judgment. ¹³, ¹⁴ We can easily tell a story about someone who hesitates about "Hesperus is Hesperus," even one in which the two tokens of "Hesperus" were associated with the same "mental file." Perhaps he is gripped by some philosophical view claiming that classical identity is incoherent. Perhaps someone he trusts dupes him into thinking that the claim is false for reasons he cannot grasp. None of this leads us to deny that there is not an a priori justification available for the thought that he resists affirming. It might be claimed, by analogy, that Fred and Twin Fred's hesitation about the proposition expressed by Fred's second judgment should provide no reason for denying the a priori of that proposition. Fred and Twin Fred have defective a priori competence, or else are suffering from a kind of noise that interferes with the exercise of that competence. ¹⁵

end p.199

I hope that at least many readers will find this response *prima facie* implausible. It contradicts what is intuitively obvious about the case, namely that from an epistemic perspective, Fred's second judgment has far more in common with the judgment that Hesperus is Phosphorus than the judgment that Hesperus is Hesperus. But let me pursue it.

A Priority

Consider a person, Oscar, who speaks both English and French. Using a hybrid of the two languages, Oscar asks himself, "Is someone bald if, and only if, that person is *chauve*?" He dithers for a while, suddenly gripped with the perhaps irrational concern that the extension of "bald" is slightly different to "*chauve*," or perhaps by misleading testimony informing him (incorrectly) that he had lost a grip on what he himself meant by "bald" and/or "*chauve*." Should we conclude from this dithering that there is no a priori resolution to the question Oscar is asking himself? We might conclude instead that the question can be resolved a priori, explaining Oscar's dithering by appeal to imperfections in his a priori competence, or noise inhibiting its exercise. Here we seem to have a case of an a priori true identity claim that is flanked by two occurrences of the same concept, but whose truth is not recognized because of some kind of a priori confusion. The confusion is not quite like that of a computational failure that inhibits a priori knowledge of some complicated mathematical identity. In the case we are interested in, the relevant claim seems to be an a priori trivial identity.

Let us call the kind of a priori confusion manifested by the "bald-*chauve*" case "Simple Identity Confusion." Our antiphysicalist suggests that our target case is an example of Simple Identity Confusion: the relevant true identity judgment can be verified a priori, and Twin Fred's inability to do so is ascribable to the kind of a priori confusion that occurs in the "bald"/"*chauve*" case.

I fear that our unclarity about the phenomenon of a priority threatens to bring the dialectic to a grinding halt. I can do little more than offer some programmatic remarks that will provide at least some reason for being dubious about the prospects for the line of resistance just sketched.

Let us begin by reflecting a little more carefully on various versions of the "bald"/"*chauve*" case. The case is underdescribed, since on some versions it is not plausible at all to suppose that the case is one of Simple Identity Confusion. Consider, for example, the following version of the case, one that emphasizes the deferential aspects of ordinary natural language: Oscar realizes that the extension of "*chauve*" is constitutively dependent on the dispositions of a particular linguistic community, while the extension of "bald" is dependent on the dispositions of a different community. Oscar is deferential enough to want the semantic value of his own use of "*chauve*" and "bald" to line up with the relevant linguistic communities. Oscar now realizes he cannot be a priori secure of an affirmative answer to the question that he is asking. No Simple Identity Confusion here.

One might suppose, however, that the direct phenomenal concepts that the antiphysicalist typically
end p.200

appeals to are not supposed to be of a deferential kind.¹⁶ So let us put to one side the point that deference can undermine a priority.

More pertinent, however, is the following case. Let us grant that Oscar is not deferential in his use of “chauve” or “bald.” To make this maximally clear, let us imagine that “bald” and “chauve” belong to two private languages that Oscar has developed for himself. Suppose further that when entertaining the hybrid question “Is something bald if and only if it is chauve?” Oscar begins to worry along the following lines: “Perhaps the dispositions that I have with regard to ‘chauve’ are slightly different from the dispositions I have with regard to ‘bald’.” Oscar “tries things out” in imagination, checking to see if his dispositions-in-imagination match up in a range of cases. But it is not clear how much a priori security this kind of process can give him. For one thing, he will likely have to do induction on a sample. But there are other kinds of worries, too: Oscar realizes that “bald” and “chauve” are context dependent. “Perhaps ‘bald’ and ‘chauve’ interact differently with certain contextual parameters in ways that I am currently unaware of,” he worries. Furthermore, he realizes that imaginative exercises performed in the cold comfort of the study could not tell him whether, for example, being placed in very hot conditions would change his dispositions to apply “chauve” and “bald” in different ways. More generally, Oscar realizes that he has no a priori security that his dispositions-in-imagination would match his actual dispositions, and he realizes that if his actual dispositions for “chauve” and “bald” come apart, that will have semantic consequences. He realizes further that even if an idealized version of himself ran through all cases in imagination, *this* kind of worry would persist.

Oscar may have heard some philosophers suggest that he try to resolve things by directly inspecting the property of baldness with his mind's eye and see if it was the same as the property expressed by “chauve.” In the face of such requests Oscar felt that he was (borrowing the words of Hilbert),

looking for something that can never be found, for there is nothing there, and everything gets lost, becomes confused and vague, and degenerates into a game of hide-and-seek. (Coffa 1991: 136)¹⁷

In short, Oscar found his question to be (1) not the kind of thing that can be answered by any kind of direct a priori inspection of the properties answering to the relevant concepts, and (2) a question to which a priori trials of his dispositions could at best give an extremely fallible and tentative answer.

Now there are cases where attempts to undermine a priori security through self-doubts about one's dispositions ring somewhat hollow. It is not as if we get a faltering a priori basis for the claim that all bachelors are unmarried by running through our dispositions in imagination and doing induction from a suitable sample.

end p.201

Here, then, we need to make explicit a further way that the “bald”/“chauve” case was underdescribed. We might suppose that there is, as a matter of deep psychology, a semantic rule requiring “bald” and “chauve” to be true of the very same things. One might try to put flesh on the bones of this picture using various kinds of deep psychological models. For example: Oscar has a single word in his language of thought, *W*, which he sometimes expresses using “chauve,” sometimes using “bald.”¹⁸ (A linguist might say that there is a single lexical entry in Oscar's language organ¹⁹: a single word in Oscar's “I-language” that gets outwardly manifested by two different words in his “E-language.”)²⁰ But I do not intend to opt for any particular model here. Let me simply assume (without argument) that there is sometimes “de jure linkage” between terms and that at least one important source of a priority is an ability to make judgments that are sensitive to such linkages (where of course sensitivity to de jure linkage is to be sharply distinguished from inferences that take claims about such linkages as premises).²¹ Here is not the place to defend the existence of semantic rules and associated de jure linkage at the level of deep psychology. I shall simply assume that such a natural kind exists.²² It seems coherent to suppose that two terms lack de jure linkage, and yet it so happens that one is disposed to apply either term in any case in which one is disposed to apply the other term (call this being “dispositionally linked”).²³ One should thus sharply distinguish between two types of cases in which one imaginatively self-examines one's dispositions to apply a term. Subcase one: as one runs through such cases, one becomes sensitive to de jure linkage between a pair of terms. Subcase two: two terms are dispositionally linked, and yet they are not de jure linked. In this case there is no de jure linkage to become sensitive to; one merely acquires some evidence of dispositional linkage through one's various imaginative exercises.

The Objection Answered

I have gestured at three models for a priori justification, one of which I suspect to be bogus.

The bogus model—as I am presenting matters—is that of Platonic Acquaintance:²⁴ one inspects a property that one stands in a cognitive acquaintance relationship with and draws cognitive insights from the process of examination. On a second model, one acquires a priori justification thanks to one's judgments' being sensitive to de jure linkage between terms (note here that it is a relation between terms rather than their denotata that is explanatorily fundamental). An imaginative game of question-and-answer would at most provide the occasion for, though not the justificatory basis of, the relevant judgments. On a third picture, an imaginative game of question-and-answer provides the fundamental evidential basis for a judgment.

Let us return to the case of Fred. I have expressed general misgivings about the Platonic Acquaintance model. Note, moreover, that such a model seems particularly unpromising as an account of a possible source of a priori justification in the particular case of Fred. Where there is a risk of qualia dancing, what could it mean to suppose that one gets a

priori justification by just staring up into Platonic Heaven using the pair of direct phenomenal concepts? I find such talk utterly unhelpful here.

Let us turn, then, to the second and third models. It does not seem plausible that there is de jure linkage between the relevant pair of phenomenal concepts, especially when one reflects that they were wrought from two separate attentional acts, and not definitionally tied to any scale. Though it may turn out that the two concepts lock on to the same property, it is unnatural to think of the case as one in which some deep semantic rule requires them to lock on to the same property. Because there is no de jure linkage to be sensitive to, the third model is of no use here for describing a source of a priori justification.

Why have I assumed that the concepts are not de jure tied to some kind of scale? Isn't it reasonable to suppose, for example, that direct phenomenal concepts get their life by de jure links to a family of background concepts that together define a phenomenal color scale? Let us recall that the neo-Russellian group proposes that the reference of a direct semantic concept is constituted by the presence of the phenomenal property to which it refers in one's phenomenal theater. Now it is important to see that this semantic thesis is altogether implausible for concepts that obtain their life via their connections to a scale. Suppose, by analogy, I am asked to form a concept of the height of a chair. Drawing on a quantitative or qualitative scale, I form a judgment: The height of the chair is "thus." In this case, direct reference semantics for the "thus" is out of the question. This is because the semantic link to the scale will trump the link to the chair in any case of conflict. Suppose "thus" stood proxy for "roughly three feet." There is no requirement at all that the chair actually be of that height. Hence it would be altogether unacceptable to suppose that the referent of "thus" was constituted by the height of the chair. Similarly, suppose I grasped a scale of pain intensities and formed a judgment, "The intensity is thus," deploying a concept from the scale. It would be out of the question to suppose that the referent of the "thus" was constituted by the actual intensity of the pain. If the reference of a direct phenomenal concept is to be

end p.203

constituted by the phenomenal type presented, then such a concept cannot get its semantic life from de jure links to a scale. And in that case, a pair of direct phenomenal concepts formed by two attentional acts cannot get de jure linked to each other via de jure links to a scale.²⁵

Let us turn to the dispositional model. Interestingly, there are severe limitations to the extent to which Twin Fred can test the sameness and difference of the denotata of the relevant phenomenal concepts by imaginative question and answer. Note that in the case of "bald" and "chauve," Oscar could provide himself with neutral descriptions of certain states of affairs ("a person with x number of hairs arranged thus and so ...") and then play out his respective dispositions with regard to "bald" and "chauve" in imagination. But in the case of Twin Fred, it does not seem that any such neutral descriptions are available. Of course, Twin Fred could instead bring to mind other phenomenal color experiences and ask whether they fall under the respective concepts. But if one thought qualia dancing was a live threat, one would be altogether hesitant about one's discriminatory capacities in new cases and would withhold judgment.

Now there are certain techniques for overcoming skeptical anxieties in the imaginative question-and-answer game. Suppose you imagine seeing a man who looks a certain way. You ask yourself whether he is bald. You could make clear to yourself that skeptical worries about perception are to be laid aside. You might ask yourself the guarded question: "Would he be bald, assuming that he is the way he looks?" That way, worries about perceptual misrepresentation could be banished as irrelevant, and you could make some judgment-in-imagination in comfort. But it is not clear what the analogous move would come to with phenomenal looks themselves (assuming one does not wish to pursue the unpromising route of postulating looks of phenomenal looks!). Suppose you are worried about qualia dancing. You form a direct phenomenal concept and then conjure up a new quale in imagination, asking yourself whether it falls under the concept previously formed. If you think that qualia dancing is a risk, then there will be no reasonable way of overcoming hesitation to apply the concept to new cases. In short, then, if you're faced with a serious prospect of qualia dancing, dispositional self-examination in imagination is no good way to test for identities of direct phenomenal concepts that are not de jure connected.

The upshot ought to be fairly clear: It seems wrong to say that in the case of Fred, an a priori justification is available for the identity claim. It doesn't exist. The antiphysicalist's response does not, then, appear to hold much promise as a defensive strategy, at least if my programmatic sketch is on the right track.

Direct Phenomenal Concepts and Antiphysicalist Arguments

I have been criticizing those who advocate blending (roughly) a Fregean approach to ordinary terms combined with a Russellian approach to direct phenomenal
end p.204

concepts. Let me now say a little about the wider implications of my critique for the debate between physicalists and antiphysicalists.

It is well known that a proposition may be metaphysically impossible and yet not a priori false. Conceivability is not a straightforward guide to metaphysical possibility. Descartes himself knew this, but he thought that under some special circumstances, conceivability was a good guide:

The rule "Whatever we can conceive of can exist" is my own, [but] it is true only so long as we are dealing with a conception which is clear and distinct. (Descartes 1647: 299)

The guiding idea, I take it, is that in a case in which one fully knows *what it is* that one is conceiving, then epistemic possibility entails metaphysical possibility. Though the language of "clear and distinct ideas" has been dropped, the basic idea is very much alive among Descartes's contemporary rationalist descendents.²⁶ Intuitively, these descendents

notice that some concepts seem "Twin Earthable" in the way that the concept of water is.

²⁷ For example, the property of evenness seems so fully present to our mind that we cannot easily imagine an epistemic counterpart who is epistemically "just like us" but who locks onto a property other than evenness. And it is natural enough to use such intuitions in support of the idea that we know what it is we are thinking about when we are thinking about evenness in a way that we may not know what it is that we are thinking about when competently exercising the concept of water. If we read "not Twin

Earthable” for “clear and distinct,” we arrive at the following version of Cartesian rationalism:

Rationalism: For thoughts that are not Twin Earthable, epistemic possibility is a guide to metaphysical possibility.

Descartes tells us, “I can achieve an easier and more evident perception of my own mind than of anything else” (Descartes 1641: 22–23). This idea finds its natural descendant in the idea that certain first-person mental concepts are not Twin Earthable. Indeed, direct phenomenal concepts seem paradigmatically not Twin Earthable. On the face of it, we know what it is we are thinking about when deploying one in such a way that there could not be an epistemic counterpart who was thinking about a different property. Such intuitions underlie Saul Kripke's famous antimaterialist argument in *Naming and Necessity*, and provide one of the central motivations for David Chalmers's two-dimensional semantics. We are quickly led to the following line of thought:

Antiphysicalism: Certain thoughts about qualia are such that if it is metaphysically possible that they are true, physicalism is not true. Those thoughts are not Twin Earthable. Moreover, they are epistemically possible (since there is no good a priori case against them). Therefore, by *Rationalism*, they are metaphysically possible. Therefore physicalism is false.

It is easy enough to see how Fred makes trouble for the kind of package offered by rationalist antiphysicalists of the contemporary stripe. Consider the thought
end p.205

“Thus is not thus,” as entertained by Fred, where the left-hand side of the identity claim expresses the phenomenal property exemplified by the left half of his phenomenal field, and the right-hand side expresses the phenomenal property exemplified by the right half. In fact the properties are identical. So given that direct phenomenal concepts are granted by all to be rigid, we seem to have a case in which a thought that is reckoned not to be Twin Earthable is metaphysically impossible and yet not a priori ruled out.²⁸ Once it is agreed that the thought is not ruled out a priori, only two options remain:

1. Reject Rationalism.
2. Concede that direct phenomenal concepts are Twin Earthable.

Making a choice here will require careful attention to how the coordinate concepts of Twin Earthability and epistemic counterpart are going to be unpacked and regimented. That is not my job here. All that needs to be observed here is that options (1) and (2) each remove a crucial premise for antiphysicalism.

Extending the Lesson

I have argued that a certain semantic package is implausible for an antiphysicalist who admits the possibility of dancing qualia. The key thought was that the thought experiments that motivate a sense-reference distinction for ordinary proper names—roughly, Hesperus-Phosphorus stories—can be replicated at the level of direct phenomenal concepts. I have shown how this thought can be motivated within the framework of a certain kind of antiphysicalism. But it is worth noticing that there may be powerful motivations for that thought even within other metaphysical frameworks. Thus the thought that Hesperus-Phosphorus phenomena arise for direct phenomenal concepts may have a quite general legitimacy. Let me make three observations in this connection. First, it might well be possible to run the dancing-qualia thought experiment even within a physicalist framework.²⁹ Suppose one is a type identity theorist about phenomenal character: each phenomenal type is identical with some neural type. Consider some such neural type *T*, which plays causal role *R*. It is arguable that the causal role of a neural state does not provide a sufficient condition for that state: a different state might have had precisely the same role. Some minimal additional assumptions will allow for the possibility of a single being in which it is sometimes *T* that plays role *R*, but sometimes a neural state that is type identical with a different phenomenal state type that plays that role. Within the framework of this kind of physicalism, our central case can be developed in essentially the same way.

end p.206

Second, I note that if indiscriminability of phenomenal character is intransitive and thus does not run in tandem with identity of phenomenal type, then a version of the Twin Fred case may be easy to contrive in the actual world.³⁰ Suppose two color chips produce phenomenally indiscriminable and yet phenomenally distinct episodes. A subject will not be able to rule out a priori the truth of the associated identity claim concerning the phenomenal characters of each episode. And yet, supposing it is flanked by rigid designators, it will be false of necessity. Thus the identity will be epistemically possible and yet metaphysically impossible.

Third, let me raise a challenge that is less theoretically laden. In the case of concepts pertaining to the empirical world, it is possible to have identity claims (about objects or properties) that are necessarily true or false without being a priori resolvable one way or another, even by someone who is fully competent (in any reasonable sense of “fully competent”) with the relevant concepts. This is because conceptual competence with empirical concepts does not bring with it a capacity to make the relevant discriminations of identity and difference. But many philosophers tacitly adopt a picture according to which, by contrast, conceptual competence with direct phenomenal concepts automatically brings with it a capacity to discriminate identity and difference of the associated properties. They allow that someone may in fact not spot the relevant identity or difference in a particular case, even a competent person. But such a case would then be explained away as a kind of competence-performance breakdown. In the case of planets, a failure to spot identity and difference in certain cases does not compel us to say that one's conceptual competence is somehow imperfect, nor that the failure issues from a

noise-involving competence-performance gap. Why should things be different in kind when it comes to phenomenal types?

Of course, one could stipulate a notion of competence such that competence in the phenomenal case brings with it a capacity to spot identity and difference—and hence that a *competent* person's failure to do so has to be ascribable to some sort of noise impeding the exercise of that capacity. But then it is far from clear that any of us is competent with phenomenal concepts in *that* sense. Pending some compelling defense of the thesis that we have some kind of capacity for perfect discrimination in the special case of direct phenomenal concepts, the neo-Russellian semantic package that has been our focus ought to look extremely suspect from the outset.

The issues can sometimes get obscured by stipulatory tactics concerning the notion of evidence. If we stipulate that one's phenomenal qualia are always part of one's evidence, then we can say that one always has evidence for identity judgments about qualia types that are manifested in one's current phenomenal theater. We can then say that Twin Fred and Fred are in different evidential situations and that Fred's evidential situation, unlike Twin Fred's, is one in which he has compelling evidence for each judgment described in the original case. Now in general we are not inclined to reckon facts that a subject is incapable of discriminating as part of the subject's evidence. Why are judgments expressed by direct phenomenal
end p.207

concepts so different? It is manifestly absurd to say that the fact that Hesperus is Phosphorus provides evidence for a belief that Hesperus is Phosphorus even if someone had no facility for discriminating that fact. And it is similarly absurd to say that the fact that one has 2,003 phenomenal red dots in one's Cartesian theater provides evidence for a belief in that fact even if someone had no facility for dot-counting. With this in mind, focus now on a qualia-dancing world in which one's left and right fields have different colors and in which an associated nonidentity claim “Thus is not thus” is true. One has no capacity whatsoever for discriminating that truth but forms the belief in the nonidentity claim just for the hell of it. Isn't it equally bad here to insist that one has evidence for one's belief?

The philosophers I am criticizing are trying to make good on a pair of ambitions: they want the propositional content of a thought token to code whether that token has an a priori justification, and they want to combine this with a direct reference semantics for certain concepts pertaining to phenomenal qualia. Now it is relatively easy to see that in the case of ordinary proper names, direct reference semantics is incompatible with the thesis that the a priority of a thought token turns on the proposition it expresses.³¹ What I have tried to show is that a similar incompatibility holds even when one restricts one's direct reference semantics to certain phenomenal concepts. The task of clarifying a priority, propositional content, and the relation between them is difficult, and much remains to be said about it. I hope I have at least clarified one corner of the terrain.

Acknowledgments

I am extremely grateful for discussions with David Chalmers. Many of the ideas and lines of thought pursued here had their roots in our conversations. I was also helped by discussions with and/or comments from Torin Alter, Brian Weatherson, Timothy Williamson, and audiences at the Universities of Nebraska and Texas at Austin.

References

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
Chalmers, D. J. (2003). The Content and Epistemology of Phenomenal Belief. In *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic: 220–72. Oxford: Oxford University Press.
Chomsky, N. (1980). *Rules and Representations*. Cambridge: Columbia University Press.
Coffa, J. (1991). *The Semantic Tradition from Kant to Carnap*. Cambridge: Cambridge University Press.
Descartes, R. (1641). *Meditations on First Philosophy*. In *The Philosophical Writings of Descartes*, Vol. 2, ed. J. Cottingham, R. Stoothoff, and D. Murdoch: 1–62. Cambridge: Cambridge University Press, 1984.
end p.208

Descartes, R. (1647). *Comments on a Certain Broadsheet*. In *The Philosophical Writings of Descartes*, Vol. 1, ed. J. Cottingham, R. Stoothoff, and D. Murdoch: 293–311. Cambridge: Cambridge University Press, 1985.
Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik* 100: 25–50. Translated as *On Sinn and Bedeutung*, in *The Frege Reader*, ed. M. Beaney: 151–71. Oxford: Blackwell, 1997.
Kripke, S. (1972). Naming and Necessity. In *The Semantics of Natural Language*, ed. G. Harman and D. Davidson. Dordrecht: Reidel. Reprinted as *Naming and Necessity*. Cambridge: Harvard University Press, 1980.
Russell, B. (1918). The Philosophy of Logical Atomism. *Monist* 28. Republished in *The Philosophy of Logical Atomism*, ed. D. Pears: 35–155. LaSalle, Ill.: Open Court, 1985.
Soames, S. (2003). *Beyond Rigidity*. Oxford: Oxford University Press.
Williamson, T. (1990). *Identity and Discrimination*. Oxford: Blackwell.
end p.209

eleven Property Dualism, Phenomenal Concepts, and the Semantic Premise

Stephen L. White

Phenomenal Concepts

The property dualism argument originated in an objection to theories that entail two theses: that mental states such as pain are identical to physical states of the brain and that mental-physical identities are a posteriori.¹ Suppose that pain is identical to C-fiber firing (CFF). Since the identity is a posteriori, a subject unaware of this fact could rationally believe what would be expressed by saying, “I am in pain, but my C-fibers are not firing.” The subject is saved from irrationality, despite believing incompatible things of the same event—the occurrence of the pain that *is* the firing of the C-fibers—by the fact that that event figures in the conjunctive belief under two distinct modes of presentation. In one, we take it, the pain (= the CFF) is given as a certain kind of brain state, and in the other it is given in the way pains are normally given when one has them. And suppose we think, in the first instance, of modes of presentation as aspects of the way we represent the world and not the world itself. (Call these *representational modes of presentation* [RMPs].) Then it seems that there must be *corresponding* features or properties of items in the world (in this case, features of the C-fiber firings)—by virtue of which the representational modes of presentation pick them out. (Call such features *nonrepresentational modes of presentation* [NMPs].) It is the burden of the property dualism argument to show that the feature or property of CFF by virtue of which it is picked out by an expression such as “my pain” (under normal circumstances) must be mental. Thus the argument purports to show that a full explanation of the a posteriori character of such mental-physical event identities presupposes a higher level mental-physical dualism (property dualism). If so, then such identity theories are incompatible with a physicalist conception of the world.

end p.210

Proponents of the so-called phenomenal concepts analysis of such mental-physical identities claim to answer this objection to physicalism (Loar 1990/97; Block, chap. 12, this volume). Such analyses purport to do full justice to the meaning and the a posteriori character of these identities in physicalist terms and their proponents have offered allegedly conclusive objections to the property dualism argument itself. In this section, I shall address these analyses directly and argue that there is no alternative here to property dualism. I shall then go on to present a version of the property dualism argument and to address the relevant objections.

Phenomenal concepts analyses entail that phenomenal concepts, such as the concept of being a pain, satisfy three conditions.

(1) *Phenomenal concepts are not equivalent to physical-functional concepts.* Hence, contrary to what is believed by analytical functionalists, for example, identities such as “pains are identical with C-fiber firings” are, as alleged, a posteriori.

(2) *Phenomenal concepts pick out their referents directly.* The relation of “pain” to pain, then, is not mediated by a mode of presentation of pain. In this it differs from the referential relation commonly thought to hold between “Hesperus” or “Phosphorus” and Venus.

The ordinary assumption is that the reference, for example, of “Hesperus” to Venus is mediated by a description such as “the first heavenly body visible in the evening.” And it is assumed that “Hesperus” picks out Venus *by virtue of* the latter's having the property expressed by the predicate contained in such a description—the property, in this case, of being the first heavenly body visible in the evening. To say that the referential relation of “pain” to pain is unmediated by a mode of presentation (in any ordinary sense) is not, of course, to say *how* phenomenal concepts *do* pick out their referents. On this score there are two distinct suggestions.

a. *Recognitional concept view*. Phenomenal concepts are type demonstratives that pick out the properties to which they refer by virtue of the successful concept user's capacity to *recognize* the relevant properties. For example, one might refer to “that taste” or to the occurrence of “that shade of red,” and one might have the capacity not only to discriminate them now but to recognize them under other (e.g., counterfactual or future) circumstances. (Loar 1990/97: 600–03)

b. *Quotation concept view*. Phenomenal concepts are sometimes alleged to pick out the properties they do by virtue of their “embedding” or “quoting” of instances of those properties. For example, the concept expressed by “this kind of pain” may actually embed an instance of the type of pain in question. (Block 2002: 396–98; Block, chap. 12, this volume)

(3) *Phenomenal concepts are not “blind.”* On the recognitional concept view, we have a (second-order) phenomenal concept of what all (first-order) phenomenal properties have in common. Alternatively, we might put this by saying that we have a recognitional concept of what is common to all the things in all the extensions of the first-order phenomenal concepts (something we might call their “phenomenality”).

Thus we have a concept of what it is by virtue of which phenomenal states differ from those picked out by (mere) self-directed recognitional concepts. And according to the recognitional concepts view, it is a confusion between phenomenal concepts and mere self-directed recognitional concepts that leads to the charge that
end p.211

applications of the former are blind (Loar 1990/97: 603–04). Those who hold the quotation concept view, of course, already have a reply to the charge of blindness. For those who hold this version of the phenomenal concept view, in deploying such concepts we actually *have* (or have something similar to) an experience of the type that the concept picks out.

The question now is whether phenomenal concepts supply an adequate explanation of the a posteriori character of mental-physical identities. There is an obvious temptation to say yes, since it is a basic tenet of the theory of phenomenal concepts that they are not equivalent to any concepts in the physical-functional domain. This point could be spelled out further by saying that the two sorts of concepts have different conceptual or inferential roles. There are no relevant entailment relations between them, and, prior to

the discovery of the a posteriori mental-physical identities, they are triggered by different sorts of experiences. For example, the experience that is ordinarily expressed by "I'm in pain" is triggered by pain as it is normally given to the subject of the experience. In contrast "my C-fibers are firing" is (for those who have the concept) triggered by (what is alleged to be) the same state as given through the experience of brain-scan devices and the like.

Plausible as this sounds, however, it must, it seems, be wrong. What is required to explain fully the a posteriori character of the mental-physical identities is not just that the concepts flanking the identity sign have different conceptual roles in this sense. What is required is an explanation of how the subject who claims sincerely not to believe such an identity takes the world to be. This is because the view that such identities have an a posteriori character entails that a subject could be fully rational while failing to believe or disbelieving them. Thus there must be a clear account of what the world would be like if it were the way that such an uninformed or misinformed (but still fully rational) subject took it to be. For suppose there were no such account, that every attempt to provide one led to incoherence. Then the truth of at least some of the mental-physical identities would be a priori, and the proponent of the phenomenal property approach to mental-physical identities would lack an explanation of their alleged a posteriori character.

Call the requirement that there be a coherent account of what the world would be like if it were the way such an uninformed or misinformed subject takes it to be the requirement of *representational coherence*. What we have just seen is that representational coherence is forced upon us by the assumption that the identities in question are a posteriori. And this requirement is stronger than Frege's constraint as it is normally understood. Schiffer formulates Frege's constraint as follows:

If x believes y to be F and also believes y not to be F , then there are distinct modes of presentation m and m' such that x believes y to be F under m and disbelieves y to be F under m' . (1978: 180)

But although this is entailed by representational coherence, it does not explicitly say that the modes of presentation should be such that representational coherence is satisfied.

To see that representational coherence is what is required, notice that there is a sense in which Frege's constraint is satisfied in the case of the person who (as we would say) believes that $27 + 17 = 44$. There is clearly a sense in which the subject

has different modes of presentation of 44 associated with the two numerical expressions.

The subject may recognize that it is twice 22, for example, under the mode of presentation associated with the right-hand expression, but not under the mode of presentation associated with the left-hand expression. Moreover, this subject satisfies all of the conditions that we earlier imagined the phenomenal property theorist offering as his or her account of the a posteriori character of the relevant mental-physical identities.

First, the two referring expressions have different conceptual roles. That is, the subject doesn't infer (or become disposed to assent to) sentences containing one from sentences containing the other. Second, he or she does not associate them with different evidential conditions (e.g., counting out 44 objects by counting to 44; counting out 44 objects by counting to 27, then counting to 17). But $27 + 17 = 44$ is *not* a posteriori. Hence what we imagined the phenomenal property theorist as offering is insufficient.

Can the phenomenal property theorist give us more by way of explanation of the a posteriori character of the mental-physical identities? In particular, can such a theorist do

justice to representational coherence? It seems not. To satisfy representational coherence in the case of such alleged identities as “pain = CFF,” we need a coherent possibility to serve as the content of the belief of the misinformed subject who disbelieves all such mental-physical identities. But such a coherent possibility is just a possible world (where the notion of possibility in question is simply noncontradiction). Suppose there is such a possible world—one describable in complete detail without contradiction—at which pain is not identical to CFF or any other physical event. Alternatively, suppose there is a possible world where, though pain is a physical event, the qualitative aspect of pain is not identical with any physical property. Or suppose there is such a world where some property of the qualitative aspect of pain, such as the property of being hurtful, is not identical with any physical property. And so on. Then it is actually the case that one of these properties is not identical to any physical property, and we have the conclusion of the property dualism argument.

Suppose, then, that there is no such world at which pain and CFF come apart. Assuming that the identity is not false, it seems that we have two possibilities. It may be that there *is* no such coherent possibility, that the identity is, contrary to what we had assumed, *a priori* and that the subject who disbelieves it is irrational. But if this is not the case, then it seems that the subject hasn't any identity or nonidentity in mind (in any ordinary sense).

This amounts to a form of eliminativism. (See the discussion of local eliminativism below.) At most, it might be said that the subject who *believes* the identity has in mind the identity of CFF with itself. But this is something he or she could have known *a priori*. The association that a subject would express by saying “CFF is this feeling,” then, would add no new fact to what the subject could have known without the experience of pain.

What a subject who had lacked the experience would gain, we would have to suppose, in coming to believe that pain = CFF would be a new set of skills. This supposition would be analogous to the Lewis-Nemirow response to the Mary example (Jackson 1986, Nemirow 1990, Lewis 1988). I have no objection to such an account. Indeed, I am sympathetic to the view that what Mary gains is a set of action capacities. I believe, however, that such an analysis will be irreducibly

end p.213

intentional and qualitative and, hence, of no use to the physicalist.² This is, in any case, not the line that the phenomenal concepts theorist takes (as Loar and Block make explicit), since such a theorist is committed to there being genuinely mental concepts in the identities in question. Thus I shall not attempt here to provide the argument for the irreducibly intentional and qualitative character of the skills in question.

Suppose, however, that the assumption behind this argument is denied. That is, suppose it is held that one's belief that pain is not identical with CFF can have genuine content, even though there is no possible world that is the way one takes the actual world to be. And assume that despite the lack of such a possible world, one can be fully rational in holding it. What, though, does the content of the identity consist in? The answer for phenomenal property theorists—both those who hold that such concepts are a matter of type demonstratives backed by a recognitional capacity and those who hold that they are type demonstratives backed by a more or less permanent access to the *demonstratum* in question—is demonstrative or direct reference. The suggestion, then, is that whatever is

going on on the “CFF side” of the identity, on the “pain side” there is simply an unmediated connection between the word “pain” and its referent. Because the referent is given directly, there is no mode of presentation under which it is given and, hence, nothing on the side of the object (no property) by virtue of which the object is given under that mode of presentation. And since it seems clear that demonstrative reference gives us genuine representational content, it seems that the question of what the content of the identity is cannot pose a problem for the physicalist and phenomenal concepts theorist.

But could the need for a possible world that would rationalize the subject who believes that he or she is in pain but that his or her C-fibers are not firing really be denied in this way? It may appear so. Suppose it is said that such a subject is in a direct or demonstrative relation to the pain (i.e., the CFF) by virtue of which it is picked out by “pain” and in another relation to it (mediated by its tendency to cause certain instrument readings and the like) by virtue of which it is picked out by “CFF.” It could then be supposed that the rationality of the subject's disbelief in the existence of the CFF is explained by the subject's ability to *imagine* the pain (i.e., the CFF) without the usual (physical/theoretical) evidence of its existence. And for those who take this line it would be the existence of possible worlds at which pain exists without the usual third-person evidence that explains the rationality of such disbelief (Boyd 1980).

This last claim, however, seems obviously false. It seems that we can imagine the pain (or the feeling of the pain, the hurtful aspect of the feeling, etc.) without the CFF itself, and not merely without its usual physical/theoretical manifestations. To see the problem for the phenomenal concepts theorist, imagine an example in which the subject has access to the pain via *two* demonstrative relations. The first we can take, as the phenomenal concepts theorist does, to be one in which the subject refers directly or demonstratively to “this pain.” The second, we can suppose, is one in which he or she refers directly or demonstratively (via brain scan
end p.214

devices or the like) to “that brain state.”³ Since it is the phenomenal concepts theorist's claim that in such cases there is direct reference and no mode of presentation, he or she has no account of the rationality of error in *this* case in which *both* references are direct. And this simply points up the fact that it is not true that demonstrative reference is accomplished without a mode of presentation of the object—a fact that Gareth Evans's (1982) and David Austin's (1990: chaps. 1–3) examples of the demonstrative versions of Frege's problem make clear. Though the mode of presentation may not be (and cannot always be; see Austin 1990) linguistic-descriptive, such modes of presentation must exist and must represent the object if we are to explain the possibility of rational error in examples of this kind. In the case of both references (“this pain,” “that brain state”) our access to the CFF is via an aspect of it—just as is our access to the ship in the Evans case in which we say “that ship” while pointing out the right-hand window to the bow and “that ship” while pointing out the left-hand window to the stern. Similarly we have our recognitional capacities by virtue of different aspects of the objects recognized. (Many people could recognize Tony Blair by virtue of his facial features under normal

circumstances, but not by virtue of any other properties—e.g., what he would look like at a costume party or made up for an amateur theatrical.) Since both references (“this pain” and “that brain state”) presuppose modes of presentation and aspects (or properties) of the referent by virtue of which these modes of presentation pick it out, the idea (on which the phenomenal concepts view is based) of ordinary objects or events being literally their own modes of presentation cannot be made to work. The mistake seems to stem from a confusion between direct reference as it is currently understood (reference unmediated by linguistic-descriptive content) and Russellian acquaintance (Russell 1918). Were our references to our own pains direct in the Russellian sense—were they analogous to the reference, according to Russell, of the so-called logically proper names (“this,” “that”) to sense data—the referents would indeed be their own modes of presentation. But direct reference in this sense is possible precisely because sense data are *thin*. That is, they have no intrinsic features besides those of which the subject is immediately aware, and hence cannot figure in true, a posteriori mental-physical identities of the kind with which we are concerned. I conclude that direct reference in the current sense cannot play the role for which the phenomenal concepts theorist has it slated. Such theorists have given us no argument against representational coherence and hence no plausible alternative to the property dualist account of rational error regarding the mental-physical identities in question. I turn, then, to the positive argument that an adequate explanation of the a posteriori character of the relevant mental-physical identities presupposes the postulation of irreducibly mentalistic entities.

end p.215

The Explanation of A Posteriori Identities

An explanation of the possibility of true a posteriori identities requires two distinct representational modes of presentation (RMPs) of the object in question. We could not, for example, represent the beliefs of a subject who (as we would say) fails to recognize that the morning star and the evening star are one and the same planet solely in terms of a singular proposition. (A singular proposition in this case would associate Venus—the planet itself—with the property of being self-identical (Schiffer 1978).) This is clear because the denial of the identity of Venus with itself would be irrational. The two RMPs in question, however, must be capable of rationalizing—that is, justifying rationally—the beliefs (as well as the intentions and actions) of a subject who believes something of the object under one mode of presentation and fails to believe it or disbelieves it under the other. And such a justification must be available at the personal level and must characterize the way the world presents itself to the subject.

We can, as we have seen, go a step further. The explanation of such an a posteriori identity requires a set of possible worlds (worlds completely describable without contradiction)⁴ that capture the content of the relevant beliefs of a subject whom we can describe as believing and as not believing or as disbelieving something of the same object. We put this by saying that what is required is a set of possible worlds that *are* the way the subject takes the actual world to be. And this is for three related reasons.

First, suppose that there were no such set of worlds. Suppose, on the contrary, that every attempt to describe a possible world that would rationalize and justify the subject's beliefs brought to light a hidden contradiction. Then, far from rationalizing the subject, we would have revealed the subject's *irrationality*. But, if there is no way to *rationaly* fail to believe or to disbelieve the identity in question, then the identity is not a posteriori but a priori, and it was the possibility of true a posteriori mental-physical identities for which we were committed to providing an explanation.

Second, as I claimed above, in the absence of a characterization of the contents of the subject's beliefs by reference to a set of possible worlds that characterize the way the subject takes the world to be, we cannot claim to have captured the *contents* of the beliefs at all. What we believe is something about the world and reflects a way it could be, could become, or could have been. This idea is reflected in the very notion of an intentional state's pointing beyond itself and in the idea of an external state of affairs' making such an intentional state true or accurate or veridical—the idea being that true or accurate intentional states are so by virtue of “truth makers” (or an appropriate counterpart in the case of accuracy).

Third, and a related point, is that the notion of a possible world (or relevant equivalent) is presupposed by the vehicle/content distinction for beliefs—that is, the distinction between what a belief represents and what does the representing. Neither the inferential role that a belief plays nor the network of connections that

end p.216

exist by virtue of the word-to-word connections in the subject's language nor the functional or physical realization of such items is sufficient to give the subject's beliefs genuine content. In the absence of connections between the subject's words and the world, such inferential or word-to-word (or concept-to-concept) connections generate either a vicious circle or an infinite regress. In either case, they provide the subject no more than an uninterpreted formal calculus. We can put the point another way by saying that the content of belief is a matter of a *condition on the world* (a way the world must be if the belief is to be accurate), not merely a *condition of the believer*.⁵

The A Priori Connection between Representational and Nonrepresentational Modes of Presentation

To say that there must be a set of possible worlds that rationalize the beliefs (and the actions and intentions) of the subject who (as we would say) doesn't believe or disbelieves the identity in question, however, raises an obvious problem. In the case of an identity such as “the morning star = the evening star,” there is no possible world with respect to which, or at which, it fails to hold for the simple reason that there is no world at which Venus is not identical with itself. What could justify what purports to be a belief to the contrary? The answer, of course, is a set of worlds at which the properties *associated* with two distinct RMPs are instantiated by distinct objects. The properties, for example, associated with “the last heavenly body visible in the morning” and “the first

heavenly body visible in the evening,” though they are actually coinstantiated by Venus, could obviously have been instantiated by different things.

What, though, is the nature of the association between the properties and the RMPs? The set of possible worlds that justify the subject's belief must be worlds that are the way the subject takes the actual world to be. Thus the connection between the RMP and its associated property could only be one that was *given* to the subject in question. Therefore it would have to be an a priori connection by virtue of the content of the RMP (for the subject)—or an a posteriori connection by virtue of the subject's empirical beliefs. And we can eliminate the latter alternative by considering subjects who lack any relevant beliefs. In the current context this means that we can consider subjects who have no empirical beliefs that would connect pain as it is experienced directly from the first-person point of view (no mediation of brain scan devices, etc.) with any internal state characterized in neurophysiological, physical, natural, or indeed objective terms. Thus, for present purposes, we can eliminate the connection(s) via empirical beliefs. We can say that

there must be an a priori connection between the subject's term (e.g., “pain”), and its associated RMP, and the property (NMP) by virtue of which that term picks out the state it does. (Call this the *a priori condition*.)

It follows, for example, that we could not answer the objection that a circular system or an infinite regress of word-to-word connections leaves us with only an uninterpreted calculus by appealing to bare causal relations. We could not, that is, answer the objection by postulating an inferential network realized in the functional system constituted by the brain and connected to the world via external causal relations. For these are relations that need not be accessible at the personal level, and the justificatory role that the RMPs and the NMPs are postulated to play requires that the justification be available to the subject in question. That is, the appeal to external causal connections is incompatible with the goal of rationalizing the subject with respect to the belief(s) under consideration.

The inadequacy of bare causal relations in this context can be seen clearly in light of the following examples. Imagine that on the subject of Jones's honesty, Smith is genuinely irrational. Smith believes in some contexts that Jones is fundamentally honest and trustworthy and in others that Jones is fundamentally dishonest and not to be trusted. Suppose also that Smith himself would have no difficulty recognizing and acknowledging his own inconsistency were his tendency (both to affirm and deny Jones's honesty) pointed out. We would say in this case that in the relevant sense, Smith has only one representational mode of presentation of Jones. That is, the same mode of presentation figures in both his beliefs that Jones is honest and that Jones is dishonest. (To use the file-keeping metaphor, he has, *at the personal level*, only one file associated with the name “Jones.”) In terms of my earlier formulation, then, there are *not* two properties associated a priori with different linguistic expressions (or different tokens of the same linguistic expression) such that those properties could have been associated with distinct objects.

This example shows, however, that the a priori condition (the requirement that the NMP be connected a priori to the corresponding RMP) is necessary for justification. For suppose that, without Smith's knowing it, he is in contact with two distinct people—Jones and his dishonest twin. And suppose that in those contexts in which Smith is disposed to

believe that Jones is honest, Smith is almost always in contact with Jones, and that in those in which he is disposed to believe the contrary, he is in causal contact with Jones's dishonest double. Would this in any way undercut the assessment of Smith as irrational on the subject of Jones's dishonesty? The answer is no. As we have seen, Smith himself would describe his beliefs as irrational were the inconsistency pointed out, and (were it pointed out) he could recognize his relevant dispositions to behave as pragmatically self-defeating by his own lights. Thus it seems clear that we cannot rationalize an otherwise irrational belief set by appealing to subpersonal functional states and external causal chains to which the subject has no access.

Moreover, just as we cannot rationalize an otherwise irrational subject by appeal to such subpersonal or external considerations, so such considerations cannot turn an otherwise rational subject into an irrational one. Consider again the subject who believes something of Venus (say, that it is hot) under the RMP
end p.218

(associated with) “the morning star” and something incompatible (say, that it is cold) under the RMP (associated with) “the evening star.” We say that such a subject is *rational* because although he or she believes incompatible things of the same planet, there are two routes to the referent consisting in two distinct RMPs and two appropriately related NMPs. Now imagine that we make the following scientific discovery.

Astrophysicists find that the two alleged properties (being close enough to the Earth to be visible and being such as to outshine all competitors in the morning, and being so constituted in the evening) are actually both explained by a single underlying property of Venus's trajectory—say the property of being *T*. And imagine that being *T* has far greater explanatory power than any of the commonsense or theoretical properties of Venus to which we currently appeal. Suppose, finally, that on the grounds that properties must pull their weight in a causal-explanatory scheme, it is concluded that there is only one property of Venus—the property of being *T*—by virtue of which each of the two RMPs picks it out.

Should we conclude that in this case there is only one way in which Venus is being conceived and that the subject is therefore irrational in believing incompatible things of it? It seems clear that we should not. For it is perfectly coherent, perfectly intelligible, that the world should have been such that the last heavenly body visible in the morning was distinct from the first heavenly body visible in the evening. This is a real possibility, even if it is not a physical possibility (given what we are assuming are the physical laws and initial conditions). And we are committed to making sense of this possibility (either to explain the rationality of the subject or simply because we are committed to claims about the consequences had the laws or initial conditions been different). Thus we are committed to the existence of properties that do not pull their weight in a causal-explanatory scheme. At least this is so if a causal-explanatory theory is one that explains the physical/causal events at the actual world.⁶ Suppose, however, that this possibility that contravenes the actual laws of physics is a real one—a way the world could have been that is fully describable without contradiction. Then it is impossible to see what could be meant by saying that a subject who believes the world *is* this way and acts

accordingly is thereby being irrational. And it seems that no empirical discovery could undermine this assessment.

Property Dualism and the Threat of Local Eliminativism

Some will likely object to the account above on grounds that rationality is misconstrued. Rationality is (they will argue) a matter of the distinctness of the subject's RMPs, not the property or properties or NMPs by virtue of which they pick out their referent. That is to say, to the extent that intentionality requires two routes to an object of which the subject believes incompatible things, the two routes are
end p.219

provided by the difference in RMPs (such as the difference between “the morning star” and “the evening star”). On this view (call it the *single property view*), the fact that both routes go through—or pick out the object by virtue of—the same property (the property of being *T*) is unproblematic. Rationality, they would say, is a matter of the a priori and the a posteriori, whereas properties and possible worlds are a matter of possibility and necessity—and there is no excuse for confusing the two. (I shall discuss the modal issues surrounding the distinctions among so-called conceptual possibility, metaphysical possibility, and strong metaphysical possibility in more detail below.)

Such a stark distinction, however, with its obvious Kripkean overtones, represents, I believe, a misunderstanding of Kripke's contribution. For consider that a priority and necessity (as they are used here) are both intimately connected with the idea of a possible world as something completely describable without contradiction. What the person believes is the case is a way that the world could (logically or conceptually) be, hence a way that the world could be that cannot be ruled out a priori. Thus it is a way that the world could be that is describable without contradiction, and therefore a way the world could be that is representable by a set of possible worlds. To suppose otherwise is, as we have seen, to confuse the vehicle of belief with the content.

Suppose we assume otherwise. Suppose it is suggested that differences in the modes of presentation required to explain the a posteriori character of the identities in question *could* consist in causal differences alone, unavailable to the subject at the personal level. To accept such a suggestion (if it does not simply stem from a confusion of vehicle and content) is to adopt a position that commits one to what I shall call *local eliminativism*. Eliminativists regarding intentionality eschew talk of content altogether in favor of an explanation of behavior in terms drawn from the natural sciences. Similarly the local eliminativist eschews such talk in particular local contexts—in this case in precisely those contexts in which an explanation of the a posteriori character of (or the possibility of rational nonbelief or disbelief in) necessary identities is required. (These contexts we might, for obvious reasons, call the Fregean contexts). Local eliminativism, however, is untenable because it is unstable. If we are committed to ascribing beliefs in such a way as to rationalize the subject (by and large), then we are committed to doing so across the board. And if we have no such commitment, then it is unclear what we could mean by the

ascription of *content*, and we should become, regarding intentionality, eliminativists pure and simple.⁷ But whatever form of eliminativism is in question, it is clear that it rules out participation in the debate over the property dualism argument. For that debate turns on the requirements of our making sense of a posteriori identities. And for a proposition to be a posteriori is simply for it to be one that one could be rational in *believing* and rational in *disbelieving* or in *failing to believe*.

The person who claims that the a posteriori character of an identity is a matter of there being distinct representational modes of presentation of the object but not

end p.220

distinct properties, then, is in the grip of an overly rigid distinction between the a posteriori and the contingent. But to say this is not to address directly the question of where the objection goes wrong. The direct answer is that the proponent of the objection has no adequate account of what the distinctness of the two RMPs consists in—that is, no account that would justify the claim that the a posteriori character of the identity had been adequately explained. What, after all, *could* the distinctness of the RMPs consist in on such an account?

Certainly an orthographic difference—e.g., a distinction that consists in a difference in the inscription type or inscription token of the RMPs—would not be adequate. Consider the distinct inscription types “Ned,” “Block,” “Ned Block,” and “Professor Block.” Since I know one and only one person named “Ned” and one and only one person named “Block” I use these terms completely interchangeably (subtleties of style and appropriateness to the social context aside). If, then, I believe what I would express by saying something of the form “Ned is *F*” and what I would express by saying something of the form “Block is not *F*,” I could not appeal to the orthographic difference in the RMPs to rebut the charge that I am irrational. Thus an orthographic difference alone cannot provide the explanation of a posteriority that the objection in question presupposes.

The problem, in this case, is, of course, that “Ned” and “Block,” though they are orthographically different, have the *same* underlying causal, functional, and inferential role. Can the proponent of the single property model simply point to *this* difference in support of the claim that distinctness of the RMPs alone provides the explanation of the a posteriori character of the identity in question? Again there is no help for the physicalist—this time because neither causal-functional role nor inferential role (nor both taken together) are sufficient to give token utterances or inscriptions genuine representational content. As we have seen, such purely internal connections alone could never give us more than an uninterpreted calculus. And to suppose that this is all there is to our system of beliefs is to fail once again to distinguish the contents from the vehicles of those beliefs or to opt for local eliminativism.

We need, therefore, in addition to the functional or inferential role of our internal representations, some connection between those representations and the world. Could we say, then, that the combination of causal-functional role, together with an external causal connection to things in the world is sufficient? (This is perhaps the picture with which the proponent of the single property view is operating.) Again, however, the answer is no, as we saw in connection with the Smith example above. For although a system of beliefs

interpreted in this way is not an uninterpreted system (in the way that a formal calculus for which we have provided no semantics is), the interpretation is not of a kind that allows us to do justice to the rationality of the subject in question. That is, we cannot, despite the difference in the functional roles of the RMPs, say how the world presents itself to the subject such that it could be rational to believe of a single object that it does and does not have some feature. We can again put the point by saying that if the proponent of the single property view is not guilty of a vehicle-content confusion, then the view collapses into local eliminativism.
end p.221

The Four-Stage Argument for the Semantic Premise

What, then, is required to explain the a posteriori nature of the relevant identities? I shall provide the answer in the form of a set of requirements that make up a four-stage sequence, and I shall call the argument for the resulting final requirements the *four-stage argument*. The first requirement stems from the two sources we have already identified: the need to explain the possibility of a posteriori identities and the need to avoid the charge of local eliminativism. We have then

Requirement 1: We must say how the world presents itself to the subject who believes incompatible or contradictory things about the same object by providing a set of possible worlds that are the way that subject takes the actual world to be.

But what would such a world look like in the case of someone who believed what would be expressed by saying, “The morning star is inhabited and the evening star is not,” given that there is no possible world at which the morning star and the evening star are distinct? The answer, as we have seen, is given by a world at which the object that has the individuating property associated with the first expression (“the morning star”) is not identical with the object that has the individuating property associated with the second (“the evening star”). This is because, as we have also seen, we need something beyond functional states to play what we might call the rationalizing role. That is, we need something more if the RMPs are to provide a complete account of the rationality of the subject in question—the rationality of disbelief in the relevant mental-physical identities presupposed by their a posteriori status. However, two causal chains to a single object would not provide a possible world that is the way the subject takes the actual world to be. Nor could we claim, as the single property theorist does, that it is the distinctness of the expressions that explains the subject's rationality. For, as we have seen, if the distinctness is to play the justificatory role for which it is slated, it must be distinctness of *content*, not merely of orthography or of internal functional or inferential role. Thus we have

Requirement 2: We must satisfy the first requirement by providing two distinct properties of the object in question which correspond to the subject's RMPs and which are such that there is a possible world at which they are instantiated by different objects.

This leaves the question open, however, as to *what the properties are* whose instantiation by different objects at some possible worlds explains the rationality of the subject's ascription of incompatible properties to the same thing. And the answer again is relatively straightforward. As we saw in connection with the Smith example, the justification of the subject's beliefs must be available to the subject at the personal level. Thus the properties that provide the justification must be associated with the subject's RMPs in the right way. Two properties of the two distinct causal chains connecting the two RMPs to the referent, if they are properties of which the subject has no inkling, and indeed no notion, would not be associated with the RMPs in the right way. Thus there must be either an a priori association by virtue of the meanings of the RMPs or an association by virtue of the subject's other

empirical beliefs. For example, if the subject believes that the last heavenly body visible in the morning is the most massive body orbiting the sun, then there would be an association of the right kind between one of the RMPs of Venus and the property of being the most massive body in a solar orbit. However, since we are free to choose a subject who has no such empirical beliefs, we have

Requirement 3: The properties (NMPs) that explain how a rational subject could fail to believe or disbelieve an a posteriori identity must be connected to the subject's RMPs a priori.

This, however, is still not sufficient. For it is plausible to suppose that there is an a priori connection between the RMP “the diameter of the Earth” and the property of being the diameter of the Earth. And it is also plausible to suppose that there is an a priori connection between an expression of the form “ n meters in diameter” (where n is replaced by some particular number) and the property of being that number of meters in diameter. (These are a priori in the sense that the association of the properties in question with the RMPs depends on no empirical or a posteriori beliefs of the subject.) Suppose, however, that the diameter of the Earth = n meters at the actual world. It seems that there are two cases: (1) If the diameter of the Earth is identified with this dimension *at the actual world* (and n meters is understood as n times the length of the *actual* meter stick), then these two properties will be coinstantiated at every possible world. There will be no possible world that justifies the belief that the subject would express by saying, “The diameter of the Earth is not n meters.” Hence, as we have seen, there will be no way of capturing the *content* of this belief, and we will be left with local eliminativism. Suppose, though, that we impose the requirement that properties should be *thin*—that is, there is nothing to such a property over and above what is understood by the subject who understands the predicate that expresses it. Then we have case (2). The thin property of being the diameter of the Earth varies from one possible world to another since one learns what the actual diameter is a posteriori and not simply by virtue of understanding the predicate “is the diameter of the Earth.” Similarly one can understand “is n meters in diameter,” perhaps by understanding that a meter is the length of the standard meter stick in Paris (for a nineteenth-century subject) without knowing what that length is. Hence the thin properties associated with each of the RMPs will be distinct—there will be possible

worlds at which they have different extensions. And this is precisely what we need if we are to find a possible world that justifies the belief of the subject who sincerely but falsely denies that the diameter of the Earth is n meters.

Requirement 1, that there be a set of possible worlds that capture the belief of the subject who denies a true a posteriori identity, thus imposes three further requirements. As we have seen, we must identify properties of the object that could have failed to be coinstantiated (Requirement 2). And as Requirement 3 dictates, these properties must be associated a priori with the subject's RMPs. We now have

Requirement 4: The properties that satisfy Requirement (3) must be thin.

If this is the case, however, we have an argument for a weakened and slightly modified version of Brian Loar's so-called Semantic Premise. According to Loar,
end p.223

antiphysicalists (such as Kripke and, evidently, proponents of the property dualism argument) are committed to the following thesis.

Semantic Premise. A statement of property identity that links conceptually independent concepts is true only if at least one concept picks out the property it refers to by connoting a contingent property of that property. (Loar 1990/97: 600)

As I have argued elsewhere (White 2003), however, this formulation is unnecessarily strong. The (allegedly contingent) property that does the work in explaining the possibility of a posteriori identities needn't be a first-order property of the referent in question. Such an explanation would work just as well if the property were second order or higher.⁸ We can, then, reformulate the principle as follows.

Weakened Semantic Premise. A statement of property identity that links conceptually independent concepts is true only if at least one concept picks out the property it refers to by connoting a contingent property of that property, a contingent property of a property of that property, or ... and so on.

But now notice that where Loar speaks of the contingency of the connection between at least one of the two properties connoted by the expressions flanking the identity sign and the referent of the expressions (or, in the weakened version, the referent, a property of the referent, a property of a property of the referent, or ... and so on). Requirements (1)–(4) turn on the contingency of the *coinstantiation* of the properties (NMPs) connoted. In line with the four-stage argument, then, we have a second reformulation of the principle:

Weakened, Modified Semantic Premise. A statement of identity that links conceptually independent concepts is true only if the expressions flanking the identity sign pick out their referents by connoting contingently coextensive properties of that referent, or contingently coextensive properties of a property of that referent, or ... (and so on).⁹ What, then, are the two properties that could serve as the routes to the referent (the event which is the pain, i.e., the CFF) for the person who doubts that any “mental” event (property, etc.) is identical with anything physical and who is perfectly rational? The only

property by virtue of which “pain” could pick out the CFF for such a subject would be something like the property of being painful or hurtful. Now consider any physical property *P*. Certainly, it seems, we could imagine the following: our being in a state that is hurtful without our being in a state that

end p.224

has *P*. But because the properties in question are thin and are connected a priori with the relevant predicates, there is no room for an illusion of contingency that is not real contingency. If so, the identity in question is false, and we have property dualism. (And if some of these conditions are not satisfied, we have no explanation of the rationality of skepticism about the relevant identities *until* we reach a level of higher order properties at which they *are* satisfied.)

Metaphysical and Conceptual Possibilities

A common objection to the property dualism argument is that the a posteriori character of mental-physical identities is a matter of conceptual possibilities or conceptually possible worlds, but that such worlds or possibilities need not be metaphysical possibilities. Thus, it is claimed, one cannot argue from the assumed a posteriori nature of the identity of, say, the property of being hurtful and some physical property to the conclusion that there is a real possible world at which they are distinct, and thus to the conclusion that they are not identical at the actual world. Indeed, such a claim can be supported by the morning star/evening star example itself. “The morning star = the evening star” is a posteriori, and it is argued that there is, therefore, a logically or conceptually possible world at which they are distinct. It is alleged, however, that there is no metaphysically possible world at which this is the case. Thus it is suggested that the property dualism argument must ultimately rest on a non sequitur.

As applied to the property dualism argument, this is misdirected. First, the property dualism argument involves no distinction between conceptually possible and metaphysically possible worlds of the sort alleged. If “pain” and the “CFF” refer rigidly, then, as we have seen, there is *no* possible world describable in complete detail and without contradiction at which pain is not identical with CFF. And it is precisely this fact that motivates the property dualist's search for the properties (or properties of properties, etc.) that are necessary for a complete and adequate explanation of the a posteriori character of what is assumed to be the *identity* of pain and CFF. There is, then, in the property dualism argument, nothing that would license a move from the a posteriori character of “pain = CFF” to the claim that there is a possible world at which they are distinct. Thus there is no illegitimate inference from something called conceptual possibility to some distinct notion of real possibility. There is one notion of a possible world used throughout—describability in complete detail without contradiction (keeping the meanings of our terms and the actual-world referents of our rigid terms fixed). And this notion is the one appropriate to the task of explaining the a posteriori character of the alleged mental-physical identities.

To their credit, the most prominent critics of so-called conceivability arguments (among which the property dualism argument is usually included) do not suppose that they involve such a crude mistake. Katalin Balog (1999) and Joseph Levine (2001), for example, distinguish “naive” conceivability arguments from arguments such as those of David Chalmers (1996) and Frank Jackson (1982, 1993, 1998)

end p.225

(and, by extension, from the property dualism argument).¹⁰ And both Balog and Levine acknowledge that the proponents of the latter have no difficulty doing justice to a posteriori necessities of the water/H₂O and morning star/evening star varieties. It would be a mistake, however, to assimilate the property dualism argument to conceivability arguments of even the sophisticated sort. (Balog does, as Levine does not, draw a clear distinction between them.) This is because the main arguments of Jackson and Chalmers turn on issues of modality.

In particular, they turn on a conception according to which modal truths are reducible to, or consist of, conceptual truths plus empirical facts about the referents of our referring expressions at the actual world. Such arguments can be countered by accepting brute modal truths that are not reducible in this way—for example, a brutally necessary connection between a subject's being in a certain physical brain state and being in a certain qualitative state. On such a view, then, there will be worlds that are possible in the sense of being completely describable without contradiction (even keeping our language and all the relevant actual-world references fixed) that are not possible metaphysically. That is, there will be coherently describable worlds that are not possible in the light of all the truths that are brutally necessary. Call the operative notion of necessity *strong metaphysical necessity*. Since there will be no inference from possibilities in my sense to strong metaphysical possibilities, does this mean that the property dualism argument begs the question against such theorists?

The answer, of course, is that the property dualism argument turns on different considerations altogether. We can see this most clearly if we simply concede (for the sake of argument) all the brute necessities that the proponent of strong metaphysical necessity desires and recognize that the property dualism argument still raises precisely the same difficulty for the physicalist. The crucial move is, as we have seen, *not* from the a posteriori nature of “pain = CFF” to the existence of possible worlds of any kind at which they are distinct. Rather, the move is from its a posteriority to the need for possible worlds that *rationalize* the subject who fails to believe or disbelieves the negation of the statement. This is just to say that if the subject is merely uninformed or misinformed and not irrational (and such a subject must be possible if the identity is a posteriori), it must be possible to give a coherent characterization of the content of the subject's belief. But this is not to say that there is a possible world in which the belief is *true*. If the subject believes what would naturally be expressed in using the sentence “The morning star is not identical with the evening star” or one of its analogues, there is, as we have seen, no such world.

Where the morning star and the evening star are concerned, the answer is straightforward. Pursuing this analogy in the case of pain and CFF, however, requires that we provide a description for the CFF as it is presented to the subject who is rational but uninformed

with respect to the identity of pain and CFF. But the only description could be something like “the state of mine that is hurtful.” And given that this must pick out the CFF by virtue of a thin property associated
end p.226

a priori with the expression, we have the property dualist conclusion. (The alternative, as we have seen, to supposing that the expression picks out the referent by virtue of a real property of the object—one that is thin and is connected to it a priori—is eliminativism regarding intentionality.)

Levine and Balog do, however, have a response that is relevant to the property dualism argument. Both hold a theory of direct reference according to which there need be no representational modes of presentation (RMPs) of the kind that the property dualism argument assumes must exist, and hence no corresponding nonrepresentational modes (NMPs). And certainly there is an important truth to what Levine calls nonascriptivism, which is the claim that when one uses a term, one need not have in mind, either “explicitly or implicitly, some description that would pick out its referent given a context” (Levine 2001: 53). But to think that direct reference in this sense is a response to the property dualism argument is to misconceive the issue, which is in essence Frege's problem, and to ignore the fact that Frege's problem arises for demonstratives as well as descriptions. As we have seen in Evans's ship example, such purely demonstrative versions of Frege's problem impose the same requirement as the standard examples—that we give an adequate account of the rationality, with respect to the relevant beliefs, of the subject who is rational but uninformed. And both Levine and Balog give at best short shrift to this requirement, which is the crux of the property dualism argument. Levine, for example, writes:

When Kripke says that what's really possible is the situation that is described had it turned out that “Water contains no hydrogen” were really true, he doesn't mean merely that we find a possible world in which those very words express a truth no matter what they mean. It's supposed to be that the situation thus picked out captures what *we really had in mind* initially by uttering the statement. So the cognitive significance of the statement must be preserved in the reinterpretation. ... In fact, it's precisely the use made of concepts and meaning by the advocates of the conceivability argument to which the [proponent of the theory Levine holds] objects. According to [such a] theorist, there is very little, if anything, like conceptual content, or cognitive significance, over and above the actual symbols of the relevant representations and their referents. (2001: 53)

Doing justice to the cognitive significance of a subject's terms, however, is not an optional feature of some solutions to the problem that motivates the property dualism argument. It *is* the problem which is defined by the requirement that we make sense of the rationality of the rational but uninformed subject. And what Balog and Levine offer—causal chains unavailable to the subject—are clearly inadequate. As we have seen in the Smith example, what rationalizes a subject with respect to particular beliefs is what is available to that subject, not a causal chain which is outside the subject's ken. Levine seems to suppose that modes of presentation (RMPs) could be unavailable to the subject because he takes “meaning” to be the object of study of an empirical science—a science, presumably, of how words function in a certain community and physical environment.

But this simply points up the fact that the issue is not one of meaning in this sense, but of *cognitive significance*—and cognitive significance for the *individual* whose rationality is in question. The crux of the problem is, after all, the task of explaining the rationality of the subject who is merely
end p.227

uninformed or misinformed regarding pain and CFF. And this is because the task is one of explaining the a posteriori character of the identity, which is simply the possibility of doubting it while being perfectly rational.

To see that the issue is one of cognitive significance in Levine's sense, imagine someone who is misinformed about the language of geometry and who believes that “right triangle” and “triangle” are synonyms. If such a subject believes what he or she would express by uttering the negation of the Pythagorean theorem, there is no irrationality of the kind involved in a straightforward belief in the negation of the theorem itself. What matters as regards rationality or irrationality are the modes of presentation available to the subject, not facts such as those about communal usage or causal chains to which he or she has no access.

Levine might respond that there is always a difference in the modes of presentation that are available to the subject in the relevant cases—namely the two distinct *expressions* (e.g., “the morning star” and “the evening star” or “pain” and “CFF”). But, as we have seen above, this proves too much. It would rationalize a completely irrational pair of beliefs—for example, beliefs that Ned Block was and was not an undergraduate physics major—simply because one was expressed using “Ned” and the other using “Block.” And this would be true even if, for the believer, the difference were merely stylistic. The objection that conceivability arguments trade on a confusion between conceptual and real possibility, then, is no threat. Either it is irrelevant to the property dualism argument, or it collapses into a direct reference claim of the kind that, in the discussion of phenomenal concepts, we have already seen reason to reject.

It is clear that in this account of the property dualism argument and of cognitive significance, the notion of thin properties does a great deal of work. Moreover, I have spoken continually of what is necessary to rationalize particular uninformed subjects in such a way that their claims to having content-laden intentional states (the uninformed beliefs) are not undermined. But I have not given a general account of what the contents of these states are. As we have seen, this is a question of what we mean when we say “what the world would be like if it were the way the subject takes it to be.” The account of the meaning of this locution and of the required notion of thinness lies in the two-dimensional semantic framework that I set out originally in “Partial Character and the Language of Thought” (1982), the necessity of which is made particularly pressing by recent objections to the property dualism argument involving special modal properties.¹¹ Thus it is to these objections that I now turn.

The Adam and Eve Objection

Consider the following objection to the Weakened, Modified Semantic Premise. The principle is intended to capture a necessary condition for there being a true a
end p.228

posteriori identity. But it could be argued that it does not. Suppose Abel is the person who originates from the union of sperm cell Adam and egg cell Eve. According to those who believe in the necessity of origins where persons are concerned, Abel could not have originated from a different sperm and egg. Thus the property of being the person who originated from sperm cell Adam is a necessary property of Abel's—that is, one that he has at every possible world at which he exists. The same is true for the property of being the person who originated from egg cell Eve. But the identity statement

1. The person who originated from sperm cell Adam = the person who originated from egg cell Eve

is clearly a posteriori. We come to believe a statement such as this on the basis of empirical investigation, and there is no difficulty in imagining a rational subject who doubts it. We have, then, apparently, a true a posteriori identity linking concepts that connote *noncontingent* properties of their common referent. Hence we seem to have a straightforward counterexample to the Semantic Premise and (as I shall assume below for the sake of argument) the Weakened Semantic Premise. The example is not, however, a counterexample to the Weakened, Modified Semantic Premise because there are worlds at which the property of being the person originating from sperm cell Adam and the property of being the person originating from Eve are not coinstantiated—for example, worlds at which sperm cell Adam is united with an egg cell other than Eve and Eve with a sperm cell other than Adam. But now consider

2. The person who originated from sperm cell Adam at the actual world = the person who originated from egg cell Eve at the actual world. ¹²

This is still a posteriori, and the properties connoted by the referring expressions are evidently coinstantiated at every possible world. Thus it seems that we have a counterexample to the Weakened, Modified Semantic Premise as well. And it would not be useful to object that the necessity of origins thesis for persons is controversial and is one that we might be tempted to deny. For whatever the truth is with regard to our own community, we can certainly imagine one in which persons are genuinely counted as the same if and only if they have the same origins in the appropriate sense.

Notice first, however, that there are two gaps in the argument against the Weakened, Modified Semantic Premise. First, there is no argument that the property of being the person who originated from sperm cell Adam is the property connoted by “the person who originated from sperm cell Adam” (and similarly for the other designating expression flanking the identity sign). Second, recall that what is being defended is the *Weakened* Modified Semantic Premise. It is not required, then, that the connoted contingent properties be properties of the referent. They might, for example, be contingently coinstantiated second-order properties (properties of

end p.229

properties of the referent), contingently coinstantiated third order properties, and so forth. Nowhere does the proponent of the objection provide an argument that all the relevant

properties have been considered. I shall refer to this latter point from time to time by saying that there must (according to the Weakened, Modified Semantic Premise) be contingently coinstantiated properties appropriately related to (connoted by) the RMPs “somewhere in the hierarchy.” In this case, however, the relevant contingent properties will turn out to be properties of the referent, Abel. Hence I shall focus almost exclusively in what follows on the first gap in the argument.

It might seem open to serious question, however, how the first alleged gap in the argument against the Weakened, Modified Semantic Premise could be exploited by a defender of the premise. How, after all, could it fail to be the case that the property connoted by “the person who originated from sperm cell Adam” is the property of being the person who originated from sperm cell Adam? Strong as the intuition supporting it may seem, however, the assumption that this is the case is extremely problematic. There is, first of all, no point in consulting our intuitions about the meaning of “connotation.” The meaning of this word is fixed by its contrast with “denotation,” but this is just to say that the connotation of a referring expression is its meaning as opposed to its referent. And to say this is simply to make explicit the extent to which the meaning of “connotation” is controversial and contested.

We can sidestep any controversy, however, about the sense of “meaning” appropriate to “connotes” by asking what the word must mean if it is to serve Loar's purpose. We are to look, that is, at what “connotes” *must* mean if it is to be plausible that the Weakened, Modified Semantic Premise provides the most charitable expression of the intuition that underpins the antiphysicalist's arguments. And doing so, I shall argue, provides an obvious and intuitive answer to the question of what connotation involves in this context. We can see what “connotation” must amount to if we consider that the Weakened, Modified Semantic Premise makes the truth of an a posteriori identity statement (one linking two conceptually independent concepts) a sufficient condition of the properties connoted by those concepts being only contingently coinstantiated. And this means that the properties connoted by the (concepts embedded in the) referring expressions must explain—indeed must constitute—their cognitive significance. For recall that a proposition is a posteriori just in case it is a possible object of belief of a perfectly rational subject and, equally, is something that a perfectly rational subject could disbelieve or doubt. And we can say that it is a sufficient condition for two coreferential referring expressions differing in their cognitive significance that a perfectly rational subject could disbelieve or doubt an identity statement linking them. And for the sake of simplicity I shall stipulate that the sufficient condition for their differing in cognitive significance is necessary as well. Thus according to this stipulation, people who doubt identity statements that are true a priori such as logical or mathematical truths—and who are, in a perfectly intuitive sense, irrational—are doubting statements linking referring expressions that do not differ in their cognitive significance. We can say, then, that a true identity statement is a posteriori if and only if it links referring expressions that differ in their cognitive significance. It follows that in order to make the best case

end p.230

for the Weakened, Modified Semantic Premise—which says (roughly) that if a true identity is a posteriori, there must be a possible world at which the connoted properties

are not coinstantiated—we must say that if the expressions flanking the identity sign differ in their cognitive significance, there must be a world at which the connoted properties are properties of different things. In other words, the properties must be at least thin enough to explain all the differences in the cognitive significance of the expressions that connote them. Of course, the properties connoted might be thinner than this. We might suppose, for example, that the property of having three sides and the property of having three angles are distinct, even though “has three sides” and “has three angles” do *not* differ in their cognitive significance (by the criterion stipulated above). However, nothing *requires* us to postulate properties that are thinner than is necessary to explain differences in cognitive significance, and doing so would prevent our identifying properties with functions from possible worlds to extensions at those worlds. Thus we can assume that the referring expressions differ in their cognitive significance if and only if the properties they connote are distinct.

A consequence of this is that what a definite description connotes cannot be assumed to be its meaning if meaning is assumed to be *broad content*. For it is arguable that in the case of the identity statement “the largest body of water on Earth = the largest body of H₂O on Earth,” the definite descriptions, in addition to being coreferential, have exactly the same broad content. Indeed, from the perspective of broad content, the suggestion that the meanings of the referring expressions are constituted (in part) by the meanings of their embedded predicates, that the meanings of those predicates are a matter of the properties they express, and that they express the same properties, is difficult to avoid. If connotation is *not* a matter of broad content, though, what account can we give of it? Must we give a complete account of so-called narrow content in order to provide an adequate interpretation of “connotes” in the context of the Weakened, Modified Semantic Premise? ¹³ The answer is no. Such an account of narrow content would have to provide an account of such topics, among others, as the contents of nonreferring expressions; the contents of beliefs of hallucinating subjects; the contents of intentional states of those who, like brains in vats, are even more radically cut off from their surroundings; and so forth. For our present purposes, however, something much more modest will do. We simply need an account of the relation between representational modes of presentation, in particular (those associated with) referring expressions, and nonrepresentational modes of presentation (properties) in cases in which the referring expressions *succeed* in picking out a referent. And we need an account according to which the NMPs explain and constitute the relations of cognitive significance among the RMPs in accordance with the four-stage argument. ¹⁴

end p.231

As we have seen, such an account has two components. There must be an a priori connection between the RMP and the corresponding NMP, and the NMP must be a thin property. Let us consider these requirements in turn. The first is relatively unproblematic. It is clear, as we have seen, that if the NMP is to provide the content of the corresponding RMP, the connection between the two must be a priori. But again, as we have seen, this is only half the requirement. Take the example of water. According to a well-known theory, what “water” refers to is H₂O. Assume that this theory is correct. That the word “water” refers to H₂O and that water *is* H₂O are empirical facts, known a posteriori. Hence

there is no a priori connection between the predicate “is water” and the property of being H_2O . The property of being H_2O , then, is *not* the nonrepresentational mode of presentation that provides the route from “water” to its referent that explains the difference (for a normal subject) in the cognitive significance of the representational modes of presentation associated with “water” and “ H_2O .”

The Definition of Thinness

Consider, though, the property of being the natural kind that falls as rain, fills the lakes and oceans, and flows from faucets *here* (or at the actual world). It seems clear that the connection between this property and “water” is a priori for normal subjects. The connection is, after all, established on the basis of a philosophical theory, not empirical research. But it seems equally clear that the property of being the natural kind that falls as rain, fills the lakes and oceans, and flows from faucets (at the actual world) has an empirical and an a posteriori *aspect*.¹⁵ We can know in advance of any empirical research (i.e., know a priori) that water *has* this property, but we cannot know (in advance of this research) what nature water has by virtue of being the natural kind that has this property. In other words, this property confers on the quantities of matter that instantiate it an *empirically discoverable essence*. Thus in terms of our earlier terminology this is a *thick* property.

We can put this point another way and say that the property is expressed by a predicate, “is the natural kind that falls as rain, fills the lakes and oceans, and flows from faucets (here),” whose associated intension (function from possible worlds to extensions) is *not* invariant with respect to *contexts of acquisition* and/or *contexts of utterance*. Acquired and uttered in this world, for example, this predicate
end p.232

determines a function from possible worlds to extensions such that at each world a substance is part of the extension if and only if it is H_2O .

Consider, though, that the same predicate (in the sense of the same linguistic expression type with the same inferential role and evidential role) was acquired and used by our functional duplicates on Twin Earth. (In this paper I shall always understand Twin Earth as an alternative possible world rather than a visitable part of this world.) In such a case, the predicate would express the intension that took possible worlds into the bodies of XYZ at those worlds. Hence the predicate is noninvariant in the sense described, and the property expressed is thick.

We have, then, three distinct conceptions of thinness where properties are concerned.

1. A property is *thin₁* if and only if it confers no empirically discoverable essence on the things in which it is instantiated.
2. A property is *thin₂* if and only if the predicate that expresses it has an intension that is invariant with respect to contexts of acquisition and/or utterance.

1. A property is *thin*₁ if and only if it confers no empirically discoverable essence on the things in which it is instantiated.
3. A property is *thin*₃ if and only if the predicate that expresses it is fully intensionalized—that is, we do not have to determine the reference of the expression at the actual world in order to determine its extension at a possible world.

(Given that the equivalence of the three definitions is perhaps not completely obvious, I shall provide the relevant arguments in the appendix.)

Let me now restate the thesis I have been defending: The Weakened, Modified Semantic Premise is correct given the two conditions already stated—that the connection between the referring expressions and the properties by virtue of which they pick out their referent must be a priori and that the properties must be thin—which we are regarding as implicit in a correct understanding of the notion of “connotation.” But why should this be the case? Won't the suspicion arise that these conditions are merely ad hoc and that even if they work in a number of cases, it is only a matter of time before we begin turning up counterexamples?

First, we should notice that the two conditions are tightly connected. Whereas the first says that the connection between the referring expressions and the properties by virtue of which they pick out their referent—or the connection between the RMPs and their corresponding NMPs—must be a priori, the second says that there must be no a posteriori component or aspect of the property. In other words, the connection must be a priori *and only a priori*. Second, the connection between the RMPs and the corresponding NMPs must be a priori and only a priori because together the RMPs and the NMPs have to explain how someone with no relevant empirical knowledge—someone whom we can imagine doubting any empirical proposition relevant to the referent—can succeed in picking it out. The connection between the NMPs and the corresponding RMPs must also be a priori and only a priori because the NMPs give content to the corresponding RMPs as demonstrated in the four-stage argument. And, as that analysis entails, they must give the kind of content that explains and constitutes the cognitive significance of the subject's representational expressions and states.

Why, though, is cognitive significance the bottom line? That it is so is built into the fundamental nature of the problem that generates the property dualism
end p.233

controversy. The problem is to explain the possibility of rational error in the case of an identity that (it is assumed) is a necessary truth. Alternatively, the problem is to explain the a posteriority of the identity. But two modes of presentation differ in their cognitive significance if and only if one can be rationally justified in believing something of a referent under one and failing to believe it, or disbelieving it, of the same referent under the other. In other words, the problem that generates the controversy just *is* the problem of explaining a difference in the cognitive significance of two modes of presentation.¹⁶ The proposal, then, is that the two conditions be treated as part of the meaning of “connotes” as it is used in the Weakened, Modified Semantic Premise. This is perfectly appropriate because “connotes” is a technical term related to “*sense*” (in Frege's sense), and cognitive significance is (arguably) the most important aspect of that concept. And,

as the foregoing suggests, this is the meaning of “connotation” that Loar needs if the Weakened, Modified Semantic Premise is to serve his purpose. Moreover, the thick/thin distinction is a stand-in for a full-blown theory of narrow content, where narrow content is the content that constitutes and explains cognitive significance and resolves the Frege puzzles. Hence it is not unmotivated to understand connotation as narrow content and to say that the Weakened, Modified Semantic Premise is correct where connotation is understood in terms of narrow rather than broad content. (This is, of course, subject to the qualification presented above that no full-blown theory of narrow content will be presented and that none is necessary in this context.) In any case, nothing turns on the identification of connotation and narrow content. We could easily restate the Weakened, Modified Semantic Premise explicitly in terms of the thick/thin distinction and avoid the use of “connotation” altogether.

Explaining the Illusion of Contingency

Even if it is clear that the “two conditions” are well motivated, we can still ask *how* the thick/thin distinction resolves the problem that (1) genuine identities are necessary, (2) the identities in question are a posteriori, and (3) there must be possible worlds that rationalize the belief of the subject who doubts ¹⁷ the identity, and (4) there must be NMPs that give the RMPs their *contents*. We can sharpen this question if we imagine a skeptical and unsympathetic objection. “You are committed,” the objector might say, “on the one hand to the idea that genuine identities are necessary—hence to the idea that there is no possible world, for example, at which pain is not the same thing as C-fiber firing (CFF). On the other hand,” the objector might continue, “You are committed to there being some connection between things on the side of rationally justified belief and RMPs and things on the side of properties,

end p.234

possible worlds, and NMPs, since you believe that the latter are necessary to give the former their content. In particular, you are committed to there being a set of possible worlds that rationalize the belief of the person who doubts or disbelieves the identity of pain and CFF. Of course, the broad/narrow or thick/thin distinctions show that these commitments are not immediately inconsistent. But you have not yet answered the question of what the possible worlds *are* that rationalize the subject's beliefs in the case of these sorts of identities.” And we can make this objection clearer if we consider a possible reply that we might make following Kripke and the objections to such a reply. The objector challenges us to satisfy the following desiderata. The first is to say which propositions they are whose *real* contingency explains the *apparent* contingency of the necessary a posteriori identities in question. The second is to say how they are related to the necessary a posteriori identities such that a fully rational subject could be justified in doubting the latter. The first desideratum amounts to the one spelled out in the four-stage argument that there be possible worlds that capture the way the subject takes the actual world to be. And the second adds to that the requirement that there be a general

characterization of the way that such worlds are related to those that constitute the truth conditions of the identities in question. Moreover, in line with the four-stage argument, these worlds that establish the genuine contingency of the proposition that justifies the subject's doubting the a posteriori identity must, in some appropriate sense, provide the content of that identity.

Consider the two replies that, on the basis of Kripke's discussion, we could make to such a challenge. First, there is the suggestion that what justifies the subject's disbelief in necessary identities of this kind is the existence of a world that is, in a qualitative sense, epistemically the same as the actual world and with respect to which a "qualitatively analogous statement" is false. For example, according to Kripke, "Heat = molecular motion" is (if, as we may assume, true) necessarily true. But its a posteriori character, apparent contingency, or the illusion of its contingency is explained by the existence of a possible world at which *the sensation of heat* is produced by something other than molecular motion. And Kripke evidently means that with regard to such a situation, the statement "The phenomenon that produces the sensation of heat is not molecular motion" is true (where by "the sensation of heat" is meant the sensation produced by heat at the actual world), even though "Heat is molecular motion" is also true at this possible world, as at every other (1972: 140–54).

Viewed as a necessary and sufficient condition (or even a necessary *or* sufficient condition) of our having an explanation of the illusion of contingency of the identity in question (or our having a rational justification of the subject who fails to believe the identity), however—and it is unclear how Kripke intended it—this suggestion is open to a number of objections. First, there seems to be no account of qualitative similarity that doesn't raise serious difficulties. If, for example, it means equivalence as regards sense data, then it relies on an analysis of perception and experience that is open to apparently conclusive objections. (I shall consider this interpretation shortly.) This is also a problem with taking the account as providing a necessary condition for our having an explanation of the apparent contingency of the identity in question or a justification of the subject who disbelieves it.

end p.235

Second, there is the Boyd objection considered above (Boyd 1980: 83–85). If the account is taken as a sufficient condition of our having the explanation and justification in question, then, as Boyd points out, it won't serve Kripke's purpose where pain and C-fiber firing are concerned. Kripke's argument is that in the case of the alleged identity of pain and C-fiber firing, the apparent possibility of our having pain without CFF and vice versa cannot be explained by citing the real possibility of our having the sensation of pain without CFF (by analogy with heat). This is because, as Kripke points out, the sensation of pain *is* pain. In other words, according to Kripke, a possible world at which we have the sensation of pain without CFF is one at which we have *pain* without CFF. Hence, according to Kripke, the attempt to explain away the apparent contingency of "pain = CFF" entails its real contingency, hence its falsity. Thus the explanation fails. According to Boyd, however, even if we assume that "pain = CFF" is true, there are (*pace* Kripke) possible worlds at which qualitative analogues of the identity are false. For there are worlds at which we have pain (and hence, by hypothesis, CFF) without any of

the sensations normally associated with CFF given *as such*—that is, as a brain state. There are, for example, obviously possible worlds at which all of our instruments give us the sensory experience associated with the absence of any CFF—they may simply not be working. (Conversely, we could have the qualitative experiences associated with CFF (given as such) without the sensation of pain, because we could have the former without CFF (i.e., pain). Thus, according to Boyd, Kripke's argument for dualism can be blocked. There is a third objection to this explanation of apparent contingency. As we saw above, the set of possible worlds that rationally justify the subject who disbelieves a necessary (but a posteriori) identity do so because they are the way the subject takes the actual world to be. They capture the *content* of the subject's belief (in the sense of the way the world presents itself to the subject), though not the truth conditions of the belief (since if pain is CFF, there is no world at which they are distinct). If this is the case, however, then pointing to a world that is qualitatively like the actual world (but at which a qualitative analogue of the identity in question is false) is not an appropriate way of rationalizing the subject and explaining the illusion of contingency. For it implies that the content of the subject's belief can be captured in sensational or sense-datum terms, and this seems clearly false. If one believes that water is not H₂O and that there can be one without the other, this seems radically different from the belief that one could have the sensations commonly associated with one without those generally associated with the other. Kripke never commits himself to an interpretation of “epistemic identity in the qualitative sense” in terms of sensations or sense data. What if we consider more liberal interpretations? Suppose, for example, that we construe epistemic equivalence in the qualitative sense as either observational equivalence in one of the many senses of “observational” or, even more vaguely, “evidential” equivalence. Would this help to address the three objections that we have just seen to what I shall call the *qualitative equivalence criterion*? Certainly it would address the first, that talk of sense data is objectionable per se. It seems less useful, however, in addressing the Boyd objection and the objection that a qualitatively specified condition cannot provide the content of the belief of the subject who doubts the pain-CFF identity. To the extent that it does address these objections, it seems to presuppose an

end p.236

instrumentalistic and phenomenalist view of science and of content that few are likely to find attractive. These objections cannot be made conclusively against the more liberal interpretation of qualitative equivalence, however, because of the vagueness of qualitative equivalence and because we cannot pursue here such topics as phenomenism, instrumentalism, and realism in the philosophy of science. Thus I shall sketch an independent problem.

Let us recall that the point of the reference to worlds qualitatively equivalent to the actual one is to explain the apparent contingency of necessary identities, and, more generally, to rationalize (i.e., rationally justify) the subject who doubts such an identity. Even more generally, it is part of the project of providing a rational justification of (intuitively) rational error wherever it occurs. Now consider a (badly misinformed) subject who believes that water is not H₂O. Following Kripke we say that his or her belief is rationalized by a possible world at which, say, the most plentiful colorless, odorless, life-

sustaining liquid is not H₂O. But now consider “colorless,” “odorless,” and “life-sustaining.” Suppose that these, like “water,” are themselves natural kind terms. Then they themselves are capable of generating exactly the same kind of apparent irrationality.

¹⁸ That is, they are capable of generating the ostensible irrationality associated with Fregean problems in which a subject doubts what he or she would express by uttering a statement that is true with regard to every possible world. If, for example, there is a discoverable empirical essence to being colorless, or odorless, or life-sustaining, then a rational subject could believe that a liquid was colorless but that it lacked the physical property with which the property of colorlessness was identical. In this case the subject would doubt a necessary truth, and we would, as yet, have no rational justification of his or her beliefs. And if we try to rule out this possibility by appealing to properties like *looking colorless*, we are back to the original phenomenalistic or sense-datum interpretation of qualitative equivalence. (And again recall that even if we don't treat so-called qualitative terms as natural kind terms, we can easily imagine rational subjects who do.)

I now turn, then, to the second way in which we might, following Kripke, say which propositions they are whose real contingency explains the apparent contingency of the necessary a posteriori identities in question. And this, of course, is to say how they are related to the latter such that a fully rational subject could be justified in doubting them. Kripke says:

In the case of identities between two rigid designators, the strategy [involving the appeal to the sameness of the subject's epistemic position, qualitatively speaking] can be approximated by a simpler one: Consider how the references of the designators are determined: if these coincide only contingently, it is this fact which gives the original statement its illusion of contingency. (1972: 150)

Unfortunately, this does not necessarily address the problems with the qualitative equivalence criterion (for example, that the representational modes of presentation that determine the references of the designators may themselves raise Fregean

problems). We can appreciate the complexity of this issue if we consider Block's example mentioned above. In the case of “The person originating from sperm cell Adam = the person originating from egg cell Eve” we have a statement that is, by hypothesis, necessary and a posteriori. And in this case, what we might call Kripke's *contingent coincidence criterion*—that if the references of the designators coincide only contingently, this gives the original statement its illusion of contingency—works to explain this fact and to rationalize the subject who disbelieves it. There are, after all, possible worlds at which “the person who originated from sperm cell Adam” and “the person who originated from egg cell Eve” are not coreferential. But now consider “The person who originated from sperm cell Adam at the actual world = the person who originated from egg cell Eve at the actual world.” This identity is a posteriori for the same reasons as the original one. But does it satisfy the contingent coincidence criterion? In fact, this version of the Adam and Eve example might seem designed to illustrate what is *right* with the criterion. Adding “at the actual world” to each of the two descriptions adds nothing to their descriptive content, since we neither identify the actual world nor pick it out from any alternatives. And to determine whether a proposition is true at the actual world or what the referent is, at the actual world, of a referring expression, we

simply determine whether the proposition is true and what the referent is. The phrase “at the actual world,” then, is best thought of, in this context, as a rigidification device. It turns a nonrigid description (or one that is not explicitly rigid) into a rigid one (anchored to the actual world) without changing its descriptive content. (We can imagine, for example, simply adding a subscript to the original description to indicate that it is to be treated as rigid. And given that it is anchored to the actual world, it seems that nothing else is required, since no descriptive content is necessary to pick out the actual world—we need neither an identifying description nor any sort of demonstrative mode of presentation of it.) But it follows that in this case, Kripke's contingent coincidence criterion is particularly easy to apply. We simply strip the rigid description of the rigidification device (imagine dropping the expression “at the actual world” or dropping the subscript) and take the nonrigid descriptions as specifying “how the references of the designators are determined.” Since, as we have seen, these do coincide only contingently, Kripke has no difficulty explaining the a posteriori character of the identity.

Though Block's particular examples pose no problem for Kripke's second criterion, there are very few cases in which the answer to the question of how the reference of the designators is determined is so straightforward. In fact, the second criterion is no real improvement over the first. In the general case, there may be no natural distinction between purely descriptive content and devices of rigidification. For there may be no identifiable language in which a purely descriptive content would find its natural expression. As we have seen, a sense-datum vocabulary generates apparently insuperable problems, and observational propositions, or evidential propositions couched in any other terms, are neither guaranteed to be purely descriptive nor to involve references that are exclusively nonrigid. It seems, then, that Kripke provides no general formula for determining which possible worlds explain the illusion of contingency of the a posteriori identities or rationally justify the beliefs of those who doubt them.

end p.238

The Dilemma for Kripke

We can, then, think of Kripke as facing a dilemma: On the one hand, Kripke can embrace a highly reductive account of the source of the illusion of contingency. That is, he can explain the illusion that water could fail to be H_2O , say, by pointing to possible worlds at which something that produces the same sensory experiences or sense data that water does (at the actual world) is not H_2O —that is, worlds at which the substance that looks colorless, has no odor, and so on is not water. This, as we have seen, is open to two objections:

1. This doesn't seem to capture what a subject means (or has to mean) in claiming that water is not H_2O . Such a subject might explicitly claim that he or she does not merely mean that something that produces the same sense data as water is not H_2O , but that *water* is not H_2O .
2. Moreover, this is open to the Boyd objection that if this is an adequate account in the case of water and H_2O , it can be applied to the case of pain and C-fiber firing to show

1. This doesn't seem to capture what a subject means (or has to mean) in claiming that water is not H_2O . Such a subject might explicitly claim that he or she does not merely mean that something that produces the same sense data as water is not H_2O , but that *water* is not H_2O .

how the illusion of contingency may be explained without denying the identity of pain and CFF. We can explain the illusion in a way exactly analogous to the heat case by pointing to the existence of worlds at which the sense data associated with CFF given as such (i.e., as a brain state) are produced by something other than pain (and hence other than CFF) and worlds at which CFF does not produce the sense data associated with CFF (given as such) at the actual world.

If, on the other hand, Kripke refuses to adopt a radically reductionistic account of the illusion of contingency, then it seems he will face an infinite regress of descriptions. If he refuses to embrace the phenomenalist strategy, then it seems he will have to explain the illusion of contingency by appeal to descriptions by virtue of which the designators in question pick out their objects. Suppose, however, that these descriptions themselves involve natural kind terms that generate further illusions of contingency—e.g., terms like “is a liquid,” understood as a natural kind term, as explained above. (And recall that even if in fact these terms are not used as natural kind terms, we can easily imagine communities in which they are.) Suppose, then, that being a liquid is itself identical to a physical property and that this identity is a posteriori. Then explaining the difference in the cognitive significance of “water” and “ H_2O ” for a normal subject in terms of the difference in the associated descriptions for that subject will lead to another set of beliefs whose obvious rationality has no explanation on Kripke's account. And it seems clear that we cannot have an infinite backward regress of such descriptive accounts. If so, then Kripke's theory will provide no general account of the rationality of error in the case of a posteriori identities.

Gap-Inducing Linguistic Devices

What is required to resolve this dilemma for Kripke is a more explicit characterization of the general account of rational error that we have already seen in outline—that is, the Weakened, Modified Semantic Premise, together with the two implicit conditions: that the relation between the RMPs and their corresponding
end p.239

NMPs must be a priori and that the properties that play the role of NMPs must be thin. The problem with what we have seen so far concerns the notion of thinness. Thin properties were originally characterized as those that conferred no hidden (or empirically discoverable) essence on the things that instantiate them. And although the application of this notion in the context of natural kind terms seems straightforward, its application in the context of the Adam and Eve example does not. To be sure, Kripke's account of the necessity of origin where persons are concerned has certain obvious analogies with the

widely accepted understanding of natural kinds. But rather than explore these analogies to expand the application of the concept of the lack of a hidden (or empirically discoverable) essence, it seems preferable that we should try to deploy a broader and more abstract conception of thinness.

It is the need for a more abstract notion of thinness that motivates the two alternatives to the characterization in terms of the absence of a discoverable essence. According to the first alternative, thin properties are those expressed by predicates whose intensions are invariant across contexts of acquisition and utterance. According to the second, thin properties are those expressed by predicates that are fully intensionalized in the following sense: given a complete description of a possible world, we can tell what the extension of the predicate is at that world without first having to answer empirical questions about the actual world. And as we have seen, properly understood, these three characterizations of thinness are equivalent.

We can, in fact, take this line of thought further. We can appeal to the notion of a predicate that is not fully intensionalized in order to define what we might call *gap-inducing linguistic devices*. Such devices take fully intensionalized predicates into predicates that are not fully intensionalized. Consider the phrase “at the actual world.” Assume (as is the case by almost anyone's lights) that “is square” is fully intensionalized. (Given a possible world PW, we can settle the question of which things are square at that world by reference to the facts at that world alone and without knowing anything about the actual world.) Contrast this with the predicate “is square at the actual world.” To know which things instantiate the property that this predicate expresses with regard to PW, we have to know what shape they have at the actual world.¹⁹ “At the actual world,” then, is a gap-inducing linguistic device.

Similarly, it would be widely accepted that proper names are such devices. Suppose we want to know with respect to a certain possible world PW' whether the following scenario obtains: that, in addition to being called “Mark Antony,” Julius Caesar did all the things associated with Mark Antony at the actual world and none of the things associated at the actual world with Caesar (i.e., Caesar played the Mark Antony role at PW'), and that Mark Antony played the Caesar role. To answer this question, we need to know who Caesar and Mark Antony are at PW'. And to do this,
end p.240

we have to determine whom “Caesar” refers to, and whom “Mark Antony” refers to, at the actual world (arguably by tracing the relevant causal chains at the actual world to their sources) and then determine what *those persons* did at PW'. In other words, we cannot settle the issue of who Caesar and Mark Antony are at PW' on the basis of what we have access to as users of the terms “Caesar” and “Mark Antony” and a qualitative description of PW'. Other plausible examples of gap-inducing devices include demonstratives and indexicals.

Something analogous might be said about natural kind terms. A plausible characterization of the content of a particular natural kind term would be something of the form
The natural kind that realizes *D* at the actual world
where *D* stands in for a description couched in terms of predicates that express the macro-level properties to which we as a community or we as individuals have access.

(Thus we might think of natural kind terms as containing implicitly the gap-inducing device “at the actual world.”) ²⁰

If this account is plausible, it suggests a general strategy for explaining rational error, a strategy that has a clear application in the Adam and Eve case. We can, as we have seen, “strip off” such apparently pure rigidification devices as “at the actual world” (devices that do not themselves contribute descriptive content to the predicate in question) while “filling in” the gaps induced by such devices as proper names and natural kind terms with descriptive content available to the subject. But, as has been implicit in what has already been said, these two strategies, even in combination, are not sufficient. Neither alone nor in combination do they deal with the problem we saw in connection with Kripke—that we cannot rule out an infinite regress of natural kind terms. We cannot rule out, for example, the possibility that every descriptive content to which we appeal to fill in a gap induced by a term like “water” (terms such as “colorless” and “wet”) will themselves be natural kind terms. And if so, there will be no descriptive, fully intensionalized vocabulary available to the subject by appeal to which all such terms could be eliminated. There is another strategy, however, by which we can abstract from any empirically discoverable essence implicit in the predicates to which our empirical situation limits us. Consider the common characterization (to which I have already alluded) of the content of a predicate (and hence of the property expressed) in terms of an intension—that is, a function from possible worlds to the extensions of that predicate at those worlds. The intuition behind the equation of such a function, the content of a predicate, and the property expressed by that predicate ²¹ is the same as the intuition that equates a proposition with a set of possible worlds. In the propositional case, the content of the proposition is whatever condition holds for all the worlds in the set. By analogy, the property expressed by the predicate is what

end p.241

the elements of all the extensions at all the possible worlds have in common. The fact that we look at the extensions of the predicates at all logically possible worlds eliminates the sort of problem raised by such predicates as “creature with a heart” and “creature with a liver.” Because these allegedly have the same extension at the actual world and yet (intuitively) differ in their contents and in the property expressed, we cannot identify their contents with their actual extensions. This problem is solved, however, by abstracting from the contingencies at the actual world. It is precisely because we can imagine a world at which creatures who in their natural, normal, and healthy state have a heart and no liver (and vice versa) that the two predicates differ in their cognitive significance and thus in their content. Thus the criterion of the identity of a property that makes it what is common to all the members of the extensions across all possible worlds is an attractive one.

The General Account of Apparent Contingency

The identification of the property connoted by a subject's linguistic expression with a set of extensions across all possible worlds does not itself provide the needed account of apparent contingency, of course. Rather, it provides the *model* for that account insofar as it involves the use of possible worlds to abstract from actual world contingencies. The difficulty, as we have seen, is that such expressions as “at the actual world,” in contexts such as “is the natural kind that realizes the ‘water role’ at the actual world,” inject actual-world contingencies not only into the *actual* extensions of predicates but into their extensions *across* possible worlds. Whereas “the natural kind that realizes the water role,” understood nonrigidly, or in a fully conceptualized way, has an extension that varies across possible worlds (it is H₂O at the actual world, XYZ at Twin Earth, etc.), “the natural kind that realizes the water role at the actual world” picks out the same natural kind at every world. Thus a contingent and empirical fact about the actual world again prevents the extension of a predicate—this time across possible worlds—from corresponding to what is, by the criterion of cognitive significance, the content of the expression in question.

There is a solution in this case, however, and it is analogous to the solution in the earlier case. We can arrive at an extension of a predicate that reflects its content if we can abstract from the actual world contingencies that break the connection between the *content* of a predicate and its extension across possible worlds. Furthermore, we have the tools at hand to do so. What I have called the *partial character* representation of content calls for a two-dimensional matrix: rows correspond to possible worlds construed as contexts of acquisition, and columns to worlds construed as contexts of evaluation.²² Now consider a fully intensionalized predicate (assume, for the sake of argument, that “is H₂O” is an example). Because the predicate is fully intensionalized, its possible-world extensions do not vary as a

end p.242

Table 11.1 Content of “is H₂O” across possible worlds

Context of Evaluation		Context of Acquisition	
AW*		TE**	
AW	H ₂ O		H ₂ O
TE	H ₂ O		H ₂ O

function of its context of acquisition (see table 11.1). That is to say, reading down each column we have the same sequence of entries as in every other column. (Equivalently, reading across each row, the entry is always the same.) Thus for fully intensionalized predicates, the matrix representation for their content reduces simply to the function from possible worlds (construed as contexts of evaluation) to extensions—that is, just the kind of function from possible worlds to extensions that has traditionally been thought to provide the contents of predicates. And for fully intensionalized predicates, this is a completely adequate characterization of their contents.

Contrast this with the case of a predicate such as “the natural kind that plays the water role at the actual world.” Acquired or uttered at the actual world, this has H₂O as its

extension at every possible world (construed as a context of evaluation). Acquired and/or uttered on Twin Earth, however, (construed as a possible world and a context of acquisition and utterance), it has as its extension XYZ at every possible world (construed as a context of evaluation) (see table 11.2).

We might put this by saying that contingencies at the context of acquisition or utterance are leveraged into necessities by our method of evaluating our subject's beliefs and utterances at possible worlds. Instead of carrying a fully intensionalized content to the various contexts of evaluation, we carry the thing that *in fact* satisfies the content (in this case a description) at the context of acquisition or utterance. We can say, then, that the function of the possible-world apparatus in bringing extensions in line with content (understood in terms of cognitive significance) that we saw earlier in connection with “is a creature with a heart” is undermined by our method of evaluating our expressions at the possible contexts of evaluation. And, of course, everything that has been said of the predicate “is the natural kind that plays

Table 11.2 Content of “the natural kind that plays the water role in the actual world” across possible worlds

Context of EvaluationContext of Acquisition

AW	TE	
AW	H ₂ O	XYZ
TE	H ₂ O	XYZ

end p.243

the water role at the actual world” has an analogue with respect to “is the person originating from sperm cell Adam at the actual world.”

But how does this treatment of possible worlds as possible contexts of acquisition help with our problem—indeed, Kripke's problem—that there may be no vocabulary available to the subject in which to express a fully intensionalized content available to the subject that has the same cognitive significance for the subject as the relevant predicate?

The answer is that there is a move open to us that is analogous to the move from taking extensions of predicates at the actual world to taking their extensions across all possible worlds in order to fix their contents. As we have seen, we can abstract from the contingency that at the actual world “is a creature with a heart” and “is a creature with a liver” have the same extensions by looking at their extensions at all possible worlds. Similarly, we can abstract from the contingency that it is H ₂ O that realizes the water role at the actual world by looking at its extensions across all possible contexts of acquisition and utterance. In other words, we explain the property that explains the possibility of rational error in a case of this kind not as the property expressed by some (set of) descriptive predicates available to the subject, but as one characterized by, reducible to, or identical with, the two- dimensional matrix that abstracts from the contingencies of the actual world not only as it is construed as a context of evaluation, but as it is construed as a context of acquisition and/or utterance.

By identifying the thin properties needed to make the Weakened, Modified Semantic Premise immune to counterexamples with the two-dimensional matrices associated with the predicates in question, we can overcome the dilemma confronting Kripke's treatment of this issue. And although Boyd's strategy for dealing with the so-called illusion of contingency in the case of pain and C-fiber firing raises a problem for Kripke, it presents no problem for the present account. Nor is there any question about whether the properties this account yields are thin enough. Keeping fixed the content (in the sense of cognitive significance) that individuals associate with the referring expressions that occur in the kinds of identities in question, we let everything else vary, compatible with consistency. If, in so doing, we produce a description of a world at which the identity, so understood, fails, then we have a rationalizing explanation and justification of the subject who calls that identity into doubt.

Does this reference to cognitive significance render the account circular? After all, one of our stated goals was to explain cognitive significance. I think the answer is no. In "Partial Character and the Language of Thought" (1982), I argued that what was held constant across possible contexts of acquisition was the functional makeup underlying the subject's use of the term in question, and this principle seems sufficient for our purposes here. For the physicalist is committed to a sense of meaning that supervenes on what is in the head—namely, the sense in which meaning just is cognitive significance. This is simply the upshot of examples we have already seen. (Recall the example of Smith, whose irrationality regarding Jones's honesty was independent of the facts regarding external causal chains precisely because the equivalence in the cognitive significance of his referring expressions was a matter of what was available to him at the personal level.) If this is the case, however, then the
end p.244

physicalist cannot object to a reductive, internal characterization of cognitive significance. (If identity or sufficient similarity of the relevant functional states is not enough, we can always add the requirement that the realizations of the functional states be "of the right physical kind.")

We can, then, appeal to the two-dimensional framework to answer the objection to Kripke regarding natural kind terms—and to provide enough of what would be provided by a positive theory of narrow content to deal with the objections to the property dualism argument and the Weakened, Modified Semantic Premise. And if we can dismiss the objections on grounds that the physicalist cannot dismiss, then the dilemma for physicalism stands: either dualism regarding sensations such as pain or properties of such sensations on the one hand, or eliminativism regarding the intentional on the other.

But note: Not only is the property dualism argument not intended to provide an account of narrow content—it is not intended to provide a positive account of any sort. It is an argument that the physicalist account in question faces an insurmountable obstacle. Thus it has been a matter of working within the physicalist framework to show that it cannot do justice to the concepts that it purports to explicate. But working within such a framework and making the minimum number of changes in order to solve isolated problems is not likely to be the best way of constructing a positive alternative. My own current preferred account of reference, for example,²³ would give pride of place not to bare causal chains,

but to irreducibly agential skills and capacities. (See the opening section of this chapter.) Such an account would require that what we keep fixed as we move from one possible context of acquisition to another are just those skills and capacities that underlie the agent's use of the relevant term. Having noted that what I have here is a negative argument and not a positive account, I can legitimately defer this discussion.

The Weakened, Modified Semantic Premise stands, then, and this account makes it clear why. The logical possibility that defines the possible worlds to which we appeal and that provides the metaphysical conclusion to the property dualism argument and the real conceivability that provides the premise are at bottom the same: describability in complete detail without contradiction. The simplicity of this connection is of course complicated significantly by the broad content that—in the form of gap-inducing linguistic devices—governs the evaluation of our utterances and beliefs at other possible worlds. It is the fundamental need to explain the possibility of rational error, however, that provides both the need for a notion of content that tracks cognitive significance and the notion that reestablishes the fundamental connection between what we can rationally conceive and what we are committed to regarding as possible in the most fundamental sense of the term.

end p.245

Appendix

Thick₁ → Thick₂

Suppose a property *P* is thick₁. Then there is some feature *F*, discoverable at the actual world, such that for any possible world, something is *P* if and only if it is *F*. But if this is how the predicate that expresses *P* works, then had the actual world been different, *P* would have been different. And this entails that the predicate does not have an intension that is invariant with respect to contexts of acquisition and/or utterance.

Thin₁ → Thin₂

Suppose there is no empirically discoverable essence. Then either (a) there is no empirical content as in, say, mathematical predicates, logical predicates, automata theoretic predicates, and so forth. Alternatively, (b) it varies from world to world. If (a), then it doesn't matter where we acquire the term; we can simply look at a possible world and tell whether something satisfies the predicate at that world.

But the same is true if (b). Consider again the description “the natural kind that falls as rain, fills the lakes and oceans, and flows from the faucets.” And suppose that the description gives the content of “water” and is understood as nonrigid. Then although at each possible world there is much to be discovered empirically about the referent of the description, there is nothing empirically discoverable that is true across all possible worlds. (All that is true across all possible worlds is that at each world where water exists it falls as rain, etc., and we know this a priori.) Hence there is no empirically discoverable *essence*. But in this case it doesn't matter where we acquire the term or where we utter it, it has the same intension—that is the same function from possible worlds to extensions.

Thin₃ → Thin₁

If a predicate has no devices of direct reference and is fully intensionalized, then there is nothing which is context sensitive. We can apply it to a possible world without having to know anything about the references of any referential devices at the actual world (besides what we know on the basis of an understanding of the language alone). Thus there is no room for a hidden or empirically discoverable essence. We could put this metaphorically by saying that there is nothing we have to “carry to another possible world,” along with the predicate, to determine its extension at that world.

Thick₃ → Thick₁

Suppose the predicate is not fully intensionalized. How, then, could we apply it to a possible world? We don't have a fully determinate content to “carry” to that world. What is the alternative? We get the necessary and sufficient conditions that we need by determining the referent at the actual world and taking the necessary and sufficient conditions for being *that thing* across possible worlds. But this gives us an empirically discoverable essence.

Acknowledgments

I am grateful to Torin Alter, Ned Block, and David Chalmers for comments and suggestions regarding earlier drafts.
end p.246

References

- Austin, D. (1990). *What's the Meaning of "This"?* Ithaca, NY: Cornell University Press.
- Balog, K. (1999). Conceivability, Possibility, and the Mind-Body Problem. *Philosophical Review* 108: 497–526. [Link](#)
- Block, N. (2002). The Harder Problem of Consciousness. *Journal of Philosophy* 99: 391–425.
- Boyd, R. (1980). Materialism without Reductionism: What Physicalism Does Not Entail. In *Readings in Philosophy of Psychology*, Vol. 1, ed. N. Block: 67–106. Cambridge: Harvard University Press.
- Brown, C. (2005) Narrow Mental Content. in *The Stanford Encyclopedia of Philosophy*, <http://plato.stanford.edu/entries/content-narrow/> .
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.
- Davidson, D. (1973). Radical Interpretation. *Dialectica* 27: 313–28. Reprinted in *Inquiries into Truth and Interpretation*: 125–170. Oxford: Clarendon, 1984.
- Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.
- Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36. [Link](#)
- Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy* 83: 291–95. [Link](#)

Jackson, F. (1993). Armchair Metaphysics. In *Philosophy in Mind*, ed. M. Michael and J. O'Leary-Hawthorne: 23–42. Dordrecht: Kluwer.

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.

Kripke, S. (1972). Naming and Necessity. In *The Semantics of Natural Language*, ed. G. Harman and D. Davidson. Dordrecht: Reidel. Reprinted as *Naming and Necessity*. Cambridge: Harvard University Press, 1980.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press. [Link](#) [OSO X-Reference](#)

Lewis, D. (1974). Radical Interpretation. *Synthese* 23: 331–44. Reprinted in Lewis, *Philosophical Papers*, Vol. 1: 108–18. New York: Oxford University Press, 1983.

[Link](#)

Lewis, D. (1983). New Work for a Theory of Universals. *Australasian Journal of Philosophy* 61: 343–77. [Link](#)

Lewis, D. (1988). What Experience Teaches. *Proceedings of the Russellian Society*. Sydney, Australia: University of Sydney. Reprinted in *Mind and Cognition*, ed. W. Lycan: 499–518. Oxford: Blackwell, 1990.

Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview. Revised version in *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzeldere: 597–616. Cambridge: MIT Press, 1997.

Nemirow, L. (1990). Physicalism and the Cognitive Role of Acquaintance. In *Mind and Cognition*, ed. W. Lycan: 490–99. Oxford: Blackwell.

Russell, B. (1918). The Philosophy of Logical Atomism. *Monist* 28. Republished in *The Philosophy of Logical Atomism*, ed. D. Pears: 35–155. LaSalle, Ill.: Open Court, 1985.

Schiffer, S. (1978). The Basis of Reference. *Erkenntnis* 13: 171–206. [Link](#)

Smart, J. J. C. (1959). Sensations and Brain Processes. *Philosophical Review* 68: 141–56. [Link](#)

White, S. (1982). Partial Character and the Language of Thought. *Pacific Philosophical Quarterly* 63: 347–65.

White, S. (1986). Curse of the Qualia. *Synthese* 68: 333–68. Reprinted as chap. 3 in *The Unity of the Self*, Cambridge: MIT Press, 1991; and in *The Nature of* [Link](#)
end p.247

Consciousness, ed. N. Block, O. Flanagan and G. Guzelere: 695–717. Cambridge: MIT Press 1997.

White, S. (1991). *The Unity of the Self*. Cambridge: MIT Press.

White, S. (1999a). Consciousness and the Problem of Perspectival Grounding. Paper presented at the Workshop on Consciousness Naturalized, Certosa de Pontignano, Siena, May 28.

White, S. (1999b). Narrow Content. In *Encyclopedia of the Cognitive Sciences*, ed. R. Wilson and F. Keil: 581–83. Cambridge: MIT Press/Bradford Books.

White, S. (1999c). Why the Property Dualism Argument Will Not Go Away.

Unpublished. Paper presented at the New York University Language and Mind Colloquium, April 4; and at the Workshop on Conceivability and Possibility, University

of Fribourg, Switzerland, December 8. Available at:
<http://www.nyu.edu/gsas/dept/philo/courses/consciousness/papers/WHYPDAW.html>
White, S. (2003) The Property Dualism Argument. Unpublished. Available at:
<http://www.nyu.edu/gsas/dept/philo/courses/consciousness/papers/White.pdf>
White, S. (2004a). Skepticism, Deflation, and the Rediscovery of the Self. *Monist* 87: 275–98.
White, S. (2004b). Subjectivity and the Agential Perspective. In *Naturalism in Question*, ed. M. De Caro and D. Macarthur: 201–27. Cambridge: Harvard University Press.
White, S. (2006). A Priori Identities and the Requirements of Rationality. In *Oxford Studies in Metaphysics II*, ed. D. Zimmerman. New York: Oxford University Press: 91–102.
end p.248

twelve Max Black's Objection to Mind-Body Identity

Ned Block

In his famous article advocating mind-body identity, J. J. C. Smart (1959) considered an objection that he thought was first put to him by Max Black. He says, “It is the most subtle of any of those I have considered, and the one which I am least confident of having satisfactorily met” (148). This argument, the “Property Dualism Argument,” as it is often called, turns on much the same issue as Frank Jackson's (1982, 1986) “Knowledge Argument,” or so I will argue. This chapter is aimed at elaborating and rebutting the Property Dualism Argument (or rather a family of property dualism arguments) and drawing some connections to the Knowledge Argument.¹ I also examine John Perry's (2001) book, which discusses both Max Black's argument and the Knowledge Argument as well as some arguments drawn from Stephen White's (1986) essay on the topic and arguments inspired by unpublished papers by White.²
I will say a bit about what the basic idea of the Property Dualism Argument is and compare it with the Knowledge Argument. Then I will discuss Perry's view of both issues. Next, I will introduce an ambiguity in the notion of mode of
end p.249

presentation and use that to give a more precise statement and rebuttal of one version of the Property Dualism Argument. In the second half of the chapter, I will use this setup to exposit and rebut another version of the Property Dualism Argument.

What Is the Property Dualism Argument?

Smart said:

Suppose we identify the Morning Star with the Evening Star. Then there must be some properties which logically imply that of being the Morning Star, and quite distinct properties which entail that of being the Evening Star. (1959: 148)

Smart goes on to apply this moral to mind-body identity, concluding that “there must be some properties (for example, that of being a yellow flash) which are logically distinct from those in the physicalist story” (148). He later characterizes the objection to physicalism as “the objection that a sensation can be identified with a brain process only if it has some phenomenal property ... whereby one-half of the identification may be, so to speak, pinned down” (149). The suggestion here is apparently that the problem of physicalism will arise for that phenomenal property even if the original mind-body identity is true. This concern motivated the “dual-aspect” theory, in which mental events are held to be identical to physical events even though those mental events are alleged to have irreducible mental properties. (See also Shaffer 1963.) Smart did not adequately distinguish between token events (e.g., this pain) and types of events (e.g., pain itself), or between token events and properties such as the property of being a pain, the property of being pain, or the property of being in pain (the first being a property of pains, the second being a property of a property, and the last being a property of persons; for purposes of this chapter, I will take types of events to be properties—any of those just mentioned will do). But later commentators have seen that the issue arises even if one starts with a mind-body property identity, even if the mind-body identity theory that is being challenged says that the property of being in pain (for example) is identical to a physical property. For the issue arises as to how that property is “pinned down,” to use Smart's phrase. If the mind body-identity says that phenomenal property Q = brain property B_{52} , then the question raised by the argument is this: Is the property by which Q is “pinned down” nonphysical or is something nonphysical required by the way it is pinned down? ³

John Perry states the argument as follows:

Even if we identify experiences with brain states, there is still the question of what makes the brain state an experience, and the experience it is; it seems like that must be an additional property the brain state has. ... There must be a property that serves as our mode of presentation of the experience as an experience... (2001: 101)

end p.250

Later, in discussing Jackson's Knowledge Argument, Perry considers the future neuroscientist Mary, who is raised in a black-and-white room (which Perry calls the Jackson Room) and learns all that anyone can learn about the scientific nature of the experience of red without ever seeing anything red. While in the room, Mary uses the term “ Q_R ” for the sensation of red, a sensation whose neurological character she knows but has never herself had. Perry says:

If told the knowledge argument, Black might say, “But then isn't there something about Q_R that Mary didn't learn in the Jackson room, that explains the difference between ‘ Q_R is Q_R ’ which she already knew in the Jackson room, and (5) [Perry's (5) is: Q_R is this subjective character], which she didn't?” There must be a new mode of presentation of that state to which “ Q_R ” refers, which is to say some additional and *apparently non-physical* aspect of that state, that she learned about only when she exited the room, that explains why (5) is *new* knowledge. (2001: 101) ⁴

On one way of understanding Perry, he uses “mode of presentation” not in the usual Fregean sense of something cognitive or semantic about a representation, but rather for a property of the represented referent. He seems to see Black's problem as arising from the

question of the physicality of the mode of presentation in that non-Fregean sense of the term. Smart speaks in the same spirit of a property that pins down one half of the identification.

The idea of the Property Dualism Argument and, I will argue, of the Knowledge Argument is that the mind-body identity approach to phenomenality fails in regard to the phenomenality that is involved in a certain kind of subjective mode of presentation (in both the Fregean and non-Fregean senses mentioned) of a phenomenal state. Even if a mind-body identity claim is true, when we look at the mode of presentation of the mental side of the identity, we are forced to accept a “double aspect” account in which unreducible phenomenal properties remain. However, don't expect a full statement of the main version of the Property Dualism Argument until nearly the halfway point. The next items on the agenda are connections to the Knowledge Argument, then Perry's solutions to both problems. After that, I take up the question of the difference between and respective roles of the Fregean and non-Fregean notions of mode of presentation. Consider a specific phenomenal property, Q, the property of feeling like the pain I am having right now. (If pain just is a type of feel, then Q is just pain.) The physicalist says, let us suppose, that Q = cortico-thalamic oscillation of

end p.251

such-and-such a kind. (I will drop the last six words.) This is an a posteriori claim. Thus the identity depends on the expressions on either side of the “=” expressing distinct concepts, that is, on their having distinct modes of presentation, for if the concepts and modes of presentation were the same, it is said, the identity would be a priori. (An ambiguity involved in this reasoning—one that derives from [surprise!] the distinction between Fregean and non-Fregean modes of presentation—will be scrutinized later in the chapter.)

“Q” in my terminology is very different from “Q_R” in Perry's terminology since “Q_R” is a term that Mary understands in the black-and-white room. By contrast, “Q” is meant (by me, even if not by Perry and Smart) as the verbal expression of a *phenomenal* concept. A phenomenal concept of the experience of red is what Mary lacked in the black-and-white room and what she gained when she went outside of it. (She also lacked a phenomenal concept of the color red, but I will not depend on that.) Why do I insist that “Q” express a phenomenal concept? Because the mind-body identity claim under consideration must be one in which the phenomenal property is referred to under a phenomenal concept of it for the Property Dualism Argument—in any of its forms—to even get off the ground. (The Knowledge Argument also depends on the use of a phenomenal concept in my sense.) Suppose that in the original identity claim we allowed any old concept of Q, such as “the property whose onset of instantiation here was at 5 p.m.” or “the property whose instantiation causes the noise ‘ouch’.” There is no special problem having to do with phenomenality for the physicalist about the cognitive significance of such properties or how such properties could pick out their referents. The modes of presentation of these properties raise no issues of the metaphysical status of phenomenality. If the original paradigm of mind-body identity were “the property whose onset of instantiation here was at 5 p.m. = cortico-thalamic oscillation,” the property in virtue of which the left-hand term presents the referent would not be a special candidate

for nonphysicality. It would be the property of being instantiated here starting at 5 p.m. . The Property Dualism Argument depends on an identity in which a *phenomenal concept* is involved on the mental side. To allow a nonphenomenal concept is to discuss an argument that has only a superficial resemblance to the Property Dualism Argument. With all this emphasis on phenomenal concepts, you may wonder what qualifies as one. A phenomenal concept is individuated with respect to fundamental uses that involve the *actual occurrence* of phenomenal properties at the time of those fundamental uses. In these fundamental uses, a simultaneous actually occurring experience is used to think about that very experience. No one could have a phenomenal concept if he could not in some way relate the concept to such fundamental uses in which the subject actually has a simultaneous instance of the phenomenal quality.

That is what I mean by a phenomenal concept, but in the rest of this chapter, I will often adopt a simplification: the fundamental uses will be taken to be all the uses of the concepts. That is, I will assume that in the exercise of a phenomenal concept, the subject actually has to have an experience. Phenomenal concepts, in this heavy-duty sense, do not really correspond to the kind of general ability that we take concepts to be individuated by. But because it is these fundamental uses that

figure in this chapter, it will make matters simpler if we usually talk about the concepts as if their only uses were the fundamental uses. The idea of these heavy-duty phenomenal concepts is that an instantiation of a phenomenal property is used in the concept to pick out a phenomenal property (a type). Of course, the experience involved in the fundamental use need not be an *additional* experience, that is, additional to the referent. A single experience can be both the object of thought and part of the way of thinking about that object. Further, one does not *have* to have an experience of red in order to think about an experience of red. One can think about the experience of red using, for example, a purely descriptive concept of it, such as “the color of ripe tomatoes.”⁵ Perry (2001, 2004a, 2004b) uses what may be a more relaxed notion of phenomenal concept, in which a phenomenal concept is a kind of mental folder that contains what he calls a “Humean idea” of the experience. He says:

Thinking of having the experience of some kind in this way is not having the experience, but it is in some uncanny way like it. Usually the same kinds of emotions attach to the thinking as to the having, although in a milder form. It is usually pleasant to anticipate or imagine having pleasant experiences, and unpleasant to anticipate or imagine having unpleasant ones, for example. (2004b: 221)

Perry's notion of a phenomenal concept is vague on the crucial point. Sure, thinking of having the experience is not just having the experience. Dogs can have experiences, but (we presume) they can't think about them. The question is, Does a phenomenal concept in Perry's sense require that the subject relate the concept to the fundamental uses I mentioned that involve an actual simultaneous experience? As I shall argue in the section on Perry below, the problem for Perry's treatment hinges on whether phenomenal concepts in his sense are phenomenal *enough* to give the Knowledge Argument and the Property Dualism Argument a fighting chance.

It is time to turn to my claim that the Knowledge Argument hinges on the same requirement of a phenomenal concept in my sense as the Property Dualism Argument. Mary is reared in a colorless environment but learns all there is to know about the

physical and functional nature of color and color vision. Yet she acquires new knowledge when she leaves the room for the first time and sees colored objects. Jackson concludes that there are facts about what it is like to see red that go beyond the physical and functional facts, and so dualism is true. From the outset, the following line of response has persuaded many critics.⁶ Mary knew about the subjective experience of red via an objective concept from neuroscience. On leaving the room, she acquires a subjective concept of the same subjective experience. In learning what it is like to see red, she does not learn about a new property. She

end p.253

knew about that property in the room under an objective concept of it, and what she learns is a new concept of that very property. One can acquire new knowledge about old properties by acquiring new concepts of them. I may know that vinegar is in the bottle and learn that acetic acid is in the bottle. In so doing, I do not learn of any new property instantiated (because the property of being vinegar just is the property of being acetic acid), and in that sense I do not learn of any new fact. I acquire new knowledge that is based on a new concept of the property that I already knew to be instantiated. When Mary acquires the new subjective concept that enables her to have new knowledge, the new knowledge acquired does not show that there are any properties beyond the physical properties. Of course, it does require that there are concepts that are not physicalistic concepts; however, that is not a form of dualism but only garden-variety conceptual pluralism: concepts of physics are also distinct from concepts of, say, economics and concepts of biology. The idea of the argument is to substitute a dualism of concepts for a dualism of properties and facts: there is a new concept but no new properties or facts in the relevant sense.

A natural rejoinder from the dualist is this. After seeing red for the first time, how does Mary “pin down” (to use Smart's obscure phrase) that old property? Or, to use an equally obscure phrase, what is Mary's “mode of presentation” of that old property?⁷ When she acquires a subjective concept of the property that she used to have only an objective concept of, *a new unreduced and unreducible subjective property* is required to “pin down” the old objective property. This is the key stage in the dialectic about Mary, and this stage of the dialectic brings in the same considerations that are at play in the Property Dualism Argument. Just to have a name for it, let us call this idea that the phenomenal concept that Mary acquires itself contains or else requires unreducible phenomenality the “metaphenomenal” move in the dialectic.⁸

The issue is sometimes put in terms of a distinction between two kinds of propositions. (See van Gulick 1993, 2006.) Coarse-grained propositions can be taken to be sets of possible worlds (or, alternatively, Russellian propositions that are *n*-tuples of objects and properties but contain no [Fregean] modes of presentation). The proposition (in this sense) that Harry Houdini escaped his bonds is the same coarse-grained proposition as the proposition that Erich Weiss escaped, in that the possible worlds in which Harry Houdini escaped are the same as the worlds in which Erich Weiss escaped, because Harry Houdini is Erich Weiss. (Alternatively,

end p.254

these are the same Russellian propositions because the proposition is the same proposition as .) Fine-grained propositions include (Fregean) modes of presentation, and so the different names determine different fine-grained propositions. When we say that Harry Houdini escaped, we express a different fine-grained proposition from the one we express when we say that Erich Weiss escaped. In these terms, the issue is: does Mary's new knowledge involve merely a new fine-grained proposition (in which case physicalism is unscathed because Mary's new knowledge does not eliminate any possibilities), or does it require a new coarse-grained proposition (as well)? It is the phenomenal mode of presentation (in the Fregean sense) of Mary's new subjective concept of a property for which she already had an objective concept that motivates the idea that she gains new coarse-grained knowledge. The metaphenomenal move is at play: the thought is that that phenomenal mode of presentation brings in something fundamentally ontological and not something on the order of (merely) a different description. The idea is that when something phenomenal is part of a (Fregean) mode of presentation, it will not do for the physicalist to say that that phenomenal item is unproblematically physical. Whether one agrees with this idea or not, if one does not recognize the idea, one misses a crucial step in the dialectic about Mary.

I said that the standard reply to Jackson's argument attempts to substitute a dualism of concepts for a dualism of properties and facts. But the dualist rejoinder that I have been describing—exploited in pretty much the same way by the Knowledge Argument and the Property Dualism Argument—is that the dualism of concepts *requires* a dualism of properties and facts.

I said that Mary acquires a subjective concept of the experience of red, whereas what she already had was an objective concept of it. However, the subjective concept she acquires is of a particular kind, namely, a phenomenal concept of the experience of red. Had she acquired an objective concept—say, the concept of the type of experience that occurred at 5 p.m. , the argument would have no plausibility. But even some subjective concepts would not do, as is the case with the concept of the type of experience that happened 5 minutes ago. This concept is subjective in that it involves the temporal location of the subject from the subject's point of view (“now”), but it is no more suitable for the Knowledge Argument than is the objective concept just mentioned. What is needed for the metaphenomenal move in the dialectic about the Knowledge Argument is that Mary acquire a mode of presentation that is either itself problematic for physicalism or that requires that the referent have a property that is problematic for physicalism. And in this regard, it is just like the Property Dualism Argument.

What Mary learns is sometimes put like this: “Oh, so *this* is what it is like to see red,” where “what it is like to see red” is a phrase she understood in the black-and-white room, and the italicized “this” is supposed to express a phenomenal concept. Because there is some doubt as to whether a demonstrative concept can really be a phenomenal concept (I'll explain the doubt below), we could put the point better by saying that what Mary learns is that P = the property of being an experience of red, where it is stipulated that “P” expresses a phenomenal concept (of a phenomenal property) and “is an experience of red” is a term Mary understood in the black-and-white room. But there is nothing special about this item of knowledge in

end p.255

the articulation of the point of the Knowledge Argument as compared with other items of knowledge that use “P.” In particular, one could imagine that one of the things that Mary learns is that $P =$ the property of being cortico-thalamic oscillation. She already knew in the room that the experience of red = cortico-thalamic oscillation (where it is understood that “the experience of red” is something she understood in the black-and-white room), but she learns that $P =$ the property of being cortico-thalamic oscillation. The proposition that $P =$ the property of being cortico-thalamic oscillation is supposed to be a new coarse-grained proposition, one that she did not know in the black-and-white room. This version of the Knowledge Argument makes the overlap with the Property Dualism Argument in the metaphenomenal move explicit: there is supposed to be something problematic about physicalism *if it is stated using a phenomenal concept*. That is, what is problematic is something about the “mode of presentation” of the phenomenal side of the identity. Both arguments can be put in the form: even if we take physicalism to be true, that supposition is undermined by the phenomenal “mode of presentation” in the knowledge or statement of it.⁹

I have used, more or less interchangeably, terms such as “pin down,” “mode of presentation,” “concept,” and “way of thinking.” But there is an ambiguity (the ambiguity between Fregean and non-Fregean readings) that must be resolved in order to focus on a precise statement of these arguments. Before I turn to that topic, however, I will give a critique of Perry's approach to Max Black, the Knowledge Argument, and modal arguments for dualism.

Perry's Treatment of the Two Arguments

Perry's (2001, 2004a, 2004b) approach to the Knowledge Argument is roughly along the lines mentioned above: Mary does something like acquiring a new subjective concept of a property that she had an objective concept of while in the black-and-white room. But Perry gives that response two new twists with two ideas: that the new concept is part of what he calls a “reflexive content” and that

end p.256

Mary need not actually acquire the new concept so long as she is appropriately sensitive to it.

Here is a quotation from Perry (2001) that gives his response both to Max Black's problem and to the Knowledge Argument:

We can now, by way of review, see how Black's dilemma is to be avoided. Let's return to our imagined physicalist discovery, as thought by Mary, attending to her sensation of a red tomato:

This i sensation = B_{52} . [where “this i ” is an internal demonstrative, and B_{52} is a brain property that she has already identified in the black-and-white room]

This is an informative identity; it involves two modes of presentation. One is the scientifically expressed property of being B_{52} , with whatever structural, locational, compositional and other scientific properties are encoded in the scientific term. This is not a neutral concept. The other is being a sensation that is attended to by Mary. This is a neutral concept; if the identity is true, it is the neutral concept of a physical property. Thus, according to the antecedent physicalist [who takes physicalism as the default view], Mary knows the brain state in two ways, as the scientifically described state and as the state that is playing a certain role in her life, the one she is having, and to which she is attending. The state has the properties that make it mental: there is something it is like to be in it and one can attend to it in the special way we have of attending to our own inner states. (2001: 205; bracketed annotations added)

If Mary's concept were "being the sensation attended to by Mary" it could not be regarded as a topic-neutral concept unless the terms "sensation" and "attend" are themselves understood in a topic-neutral manner. (Ryle introduced the term "topic-neutral" for expressions that indicate nothing about the subject matter. Smart offered topic-neutral analyses of mental terms that were supposed to entail neither that the property is physical nor that it is nonphysical. But it is clear that mentalistic terminology was supposed to be precluded, for otherwise no topic-neutral analyses would be needed—the terms would already have been topic-neutral.)

If Mary's concept is topic-neutral, it is not a phenomenal concept in the sense required by the Property Dualism Argument. Although Perry rejects the "deflationist" view that phenomenal concepts can be analyzed a priori in nonphenomenal terms (as Smart advocated), his approach to arguments for dualism is to appeal to topic-neutral demonstrative/recognition concepts as surrogates for phenomenal concepts. To explain what he has in mind, we need to introduce what he calls "reflexive content."

Propositional attitudes have "subject matter" contents that concern the properties and objects the attitudes are about. The subject matter content of your belief that the morning star rises could be taken to be the Russellian proposition. But there are other thoughts that have the same subject matter content and have the same truth condition: for example, that the heavenly object which you are now thinking of is in the extension of the property that is the object of your concept of rising. This thought has the same subject matter content but a different reflexive content. ("Reflexive" is meant to indicate that what is being brought in has to do with the way thought and language fit onto the world or might fit onto the world.) The subject matter content of the claim that $this_i$ (where " $this_i$ " is an internal demonstrative) = B_{52} , if physicalism is right, is the same as that $this_i = this_i$ or that $B_{52} = B_{52}$.

end p.257

Perry's intriguing idea is that my belief can have reflexive contents, the concepts of which are not concepts that I actually have (or even if I have them, those concepts are not ones that I am exercising in using demonstrative or recognition concepts that have those reflexive contents). However, he argues persuasively that these concepts may be psychologically relevant nonetheless if the subject is "attuned" to the concepts in reasoning and deciding. Attunement is a doxastic attitude that can have contents that are not contents of anything the subject believes or has concepts of. For example, I can be attuned to a difference in the world that makes a perceptual difference without

conceptualizing the difference in the world. Perry's view is that our intuitions about contents often involve reflexive contents that we are attuned to rather than subject matter contents that we explicitly entertain.

Perry's solution to Max Black's problem and his reply to Jackson is to focus on a topic-neutral version of what Mary learns. I am not sure whether it is just the demonstrative/recognition concept ("this_i") that is topic-neutral, or whether the reflexive content of it is also supposed to be topic-neutral. But both proposals evade the Max Black problem without solving it. In the passage quoted earlier, he says that what Mary learns can be put in terms of "This_i sensation is brain state B₅₂," where "this_i" is a topic-neutral internal demonstrative/recognition concept. If the suggestion is that Mary acquires the belief that this_i is brain state B₅₂, the problem is that the topic-neutral concept involved in this belief is not a phenomenal concept, so the real force of the Knowledge Argument (and Max Black's argument) is just ignored. However, it seems that Perry's suggestion is that Mary comes to be *attuned* to the relevant reflexive content instead of coming to believe it. He thinks that what Mary learns can be expressed in terms of something she is attuned to and that Max Black's problem can be solved by appealing to this attunement to the same content. That is, in using demonstrative and recognition concepts in the thought "This_i sensation = B₅₂," Mary becomes attuned to a reflexive content like "the sensation Mary is attending to is the scientifically described state" without explicitly exercising those concepts.

But does substituting attunement for belief avoid the objection I made that Perry is neglecting the phenomenal concepts that give the argument a chance? Does attunement help in formulating a response to the Mary and Max Black arguments that takes account of the metaphenomenal move in the Mary dialectic? I think not.

Distinguish between two versions of Jackson's "Mary." Sophisticated Mary acquires a genuine phenomenal concept when she sees red for the first time. Naïve Mary is much less intellectual than Sophisticated Mary. Naïve Mary does not acquire a phenomenal concept when she sees red for the first time (just as a pigeon, we presume, would not acquire a new concept on seeing red for the first time), nor does she acquire an explicit topic-neutral concept, but she is nonetheless *attuned* to a certain topic-neutral nonphenomenal concept such as that of "The sensation I am now attending to is the brain state I wrote my thesis on earlier." In addition, we might suppose (although Perry does not mention such a thing) that Naïve Mary is also attuned to a genuine phenomenal concept of a color even though she does not actually acquire such a concept.

As I mentioned earlier, there is a well-known solution to the Mary problem that takes Mary as Sophisticated Mary. What Sophisticated Mary learns is a
end p.258

phenomenal concept of a physical property that she already had a physical concept of in the black-and-white room. Any solution to the Mary problem in terms of Naïve Mary is easily countered by a Jacksonian opponent who shifts the thought experiment from Naïve to Sophisticated Mary. Consider this dialectic. Perry offers his solution. The Jacksonian opponent says: "OK, maybe that avoids the problem of Naïve Mary, but the argument for dualism is revived if we consider a version of the thought experiment involving Sophisticated Mary, that is, a version of the thought experiment in which Mary actually

acquires the phenomenal concept instead of merely being attuned to it (or attuned to a topic-neutral surrogate of it). What Sophisticated Mary learns is a content that contains a genuine phenomenal concept. And that content was not available to her in the room. What she acquires is phenomenal knowledge (involving a phenomenal concept), knowledge that is not deducible from the physicalistic knowledge she had in the black-and-white room. So dualism is true.”

Indeed, it is this explicit phenomenal concept that makes it at least somewhat plausible that what Mary acquires is a new coarse-grained belief as well as a new fine-grained belief. Perry cannot reply to *this* version of the thought experiment (involving Sophisticated Mary) by appealing to the *other* one (involving Naïve Mary). And the thought experiment involving Sophisticated Mary is not avoided by appeal to attunement to a topic-neutral concept or even to a phenomenal concept.

As I indicated earlier, the crucial point in the dialectic about Mary is this: The dualist says, “The concept that Mary acquires (or acquires an attunement to) has a mode of presentation that involves or requires unreducible phenomenality.” If Perry appeals to the idea that the concept is topic-neutral or has a topic-neutral reflexive content, the dualist can reasonably say, “But that isn't the concept I was talking about; I was talking about a genuinely phenomenal concept.”¹⁰

Let us now turn to Perry's solution to the Max Black problem. Although the Max Black problem is mentioned a number of times in the book, Perry's solution is expressed briefly in what I quoted above. He clearly intends it to be a by-product of his solutions to the other problems. I take it that that solution is the same as the solution to the Mary problem, namely that the problem posed by the alleged nonphysical nature of the mode of presentation of the phenomenal side of a mind-body

end p.259

identity or what is required by that mode of presentation can be avoided by thinking of what Mary learns in terms of a demonstrative/recognitional topic-neutral concept that—perhaps—has a topic-neutral reflexive content. The proponent of the Max Black argument (the property dualist) is concerned that in the mind-body identity claim “ $P = B_{52}$,” where “ P ” expresses a phenomenal concept, the phenomenal mode of presentation of P undermines the reductionist claim that $P = B_{52}$. Someone who advocates this claim—and who, like Perry, rejects deflationist analyses of phenomenal concepts—is certainly not going to be satisfied by being told that the content that Mary is attuned to is topic-neutral. The property dualist will say: “So what? My concern was that the mode of presentation of P introduces an unreducible phenomenality; whether Perry's topic-neutral content is something we believe or are merely attuned to is not relevant.” And even if what Mary is attuned to is a reflexive content that contains a genuine phenomenal concept, that also evades the issue without solving it, since the dualist can reasonably say that it is the actual phenomenal concept on which the argument for dualism is based. Perry also applies his apparatus to the modal arguments for dualism such as Kripke's and Chalmers's. Why do we have the illusion that “This ι sensation = B_{52} ” is contingent, given that (according to physicalism) it is a metaphysically necessary truth? Perry's answer is that the necessary identity has some *contingent* reflexive contents, such as: that the subjective character of red objects appears like so-and-so on an autocerebroscope, is

called “B₅₂,” and is what I was referring to in my journal articles. The illusion of contingency comes from these reflexive contents. Here, the metaphenomenal move I mentioned earlier has no role to play. I think Perry's point has considerable force. However, the dualist can respond to Perry by saying, “Look, I can identify the brain state by its *essential properties* and still wonder whether I could have that brain state (so identified) without *this phenomenal property*.” A version of this argument will be explored later in the chapter.

Though I agree with Perry on many things about phenomenality and find his book, with its notion of attunement to reflexive concepts, insightful and useful, there is one key item from which all our disagreements stem. He does not recognize the need for, or rather he is vague about the need for, a kind of phenomenal concept that itself requires fundamental uses that are actually experiential. When saying what it is that Mary learns, he says: “This new knowledge is a case of recognitional or identificational knowledge. ... We cannot identify what is new about it with subject-matter contents; we can with reflexive contents” (2004: 147). The physicalist will agree that what Mary learns is not a *new* subject matter content (in the sense explained earlier). But the problem is that it is unclear whether the recognitional or identificational concepts that Perry has in mind have the phenomenality required to avoid begging the question against the advocate of Max Black's argument. When he proposes to explain away the intuitions that motivate the Max Black argument and the Knowledge Argument by appeal to a topic-neutral concept, he loses touch with what I called the metaphenomenal move and, with it, the intuitive basis of these arguments in phenomenal concepts, or so it seems to me.

The reader may have noticed that there has still not been an explicit statement of the Property Dualism Argument. I have postponed the really difficult and controversial part of the discussion, the explanation of an ambiguity in “mode of presentation,” and I turn to that now.

end p.260

Modes of Presentation

The “mode of presentation” of a term is often supposed to be whatever it is on the basis of which the term picks out its referent. The phrase is also used to mean the cognitive significance of a term, which is often glossed as whatever it is about the terms involved that explains how true identities can be informative. (Why is it informative that Tony Curtis = Bernie Schwartz but not that Tony Curtis = Tony Curtis?) However, it is not plausible that these two functions converge on the same entity, as noted in Tyler Burge (1977) and Alex Byrne and Jim Pryor (2006).¹¹

I believe that these two functions or roles are not satisfied by the same entity, and so one could speak of an ambiguity in “mode of presentation.” However, perhaps confusingly, the Property Dualism Argument depends on a quite different ambiguity in “mode of presentation.”¹² I will distinguish between the cognitive mode of presentation (CMoP) and the metaphysical mode of presentation (MMoP). The CMoP is the Fregean mode of presentation mentioned earlier, a constellation of mental (cognitive or experiential) or semantic features of a term or mental representation that plays a role in determining its

reference or, alternatively but not equivalently, constitutes the basis of explanation of how true identities can be informative (and how rational disagreement is possible—I will take the task of explaining informativeness and rational disagreement to be the same, using “cognitive significance” for both. I will also tend to simplify, using “cognitive,” to describe the relevant constellation of features. Since semantic and experiential differences make a cognitive difference, they don't need to be mentioned separately.). The importantly different, non-Fregean, and less familiar mode of presentation, the MMoP, is a property of the referent. There are different notions of MMoP corresponding to different notions of CMoP. Thus if the defining feature of the CMoP is taken to be its role in determining reference, then the MMoP is the property of the referent in virtue of which the CMoP plays this role in determining reference. If the defining feature of the CMoP is taken to be explaining cognitive significance, then the MMoP is the property of the referent in virtue of which cognitive significance is to be explained. For example, suppose, temporarily, that we accept a descriptive theory of the meaning of names. On this sort of view, the CMoP of “Hesperus” might be taken to be cognitive features of “the morning star.” “The morning star” picks out its referent by virtue of the referent's property of rising in the morning rather than its

end p.261

property of being covered with clouds or having a surface temperature of 847 degrees Fahrenheit. The property of the referent of rising in the morning is the MMoP. (And this would be reasonable for both purposes: explaining cognitive significance and determining the referent.) The CMoP is much more in the ballpark of what philosophers have tended to take modes of presentation to be, and the various versions of what a CMoP might be are also good candidates, as good as any, for what a concept might be. The MMoP is less often thought of as a mode of presentation—perhaps the most salient example is certain treatments of the causal theory of reference in which a causal relation to the referent is thought of as a mode of presentation (Devitt 1981).

In the passage quoted earlier from Perry's statement of Max Black's argument, Perry seemed often to be talking about the MMoP. For example, he says: “Even if we identify experiences with brain states, there is still the question of what makes the brain state an experience, and the experience it is; it seems like that must be an additional *property* the brain state has. ... There must be a *property* that serves as our mode of presentation of the experience as an experience” (2001: 101, italics added). Here he seems to be talking about the MMoP of the brain state (the brain state being the experience, if physicalism is right). When he says what Max Black would say about what Mary learns, he says: “ ‘But then isn't there something about Q_R that Mary didn't learn in the Jackson room, that explains the difference between “ Q_R is Q_R ” which she already knew in the Jackson room, and (5) [(5) is: Q_R is this subjective character], which she didn't?’ There must be a new mode of presentation of that state to which ‘ Q_R ’ refers, which is to say some additional and *apparently nonphysical aspect* of that state, that she learned about only when she exited the room, that explains why (5) is *new knowledge*” (2001: 101, italics added). Again, “aspect” means property, a property of the state. So it appears that in Perry's rendition, a mode of presentation is an MMoP. However, his solution to Max Black's problem focuses on the idea that the concept that Mary acquires or acquires

sensitivity to is topic-neutral, and that makes it look as if the issue in the Property Dualism Argument is centered on the CMoP. He says, speaking of a mind-body identity: This is an informative identity; it involves two modes of presentation. One is the scientifically expressed *property* of being B₅₂, with whatever structural, locational, compositional and other scientific properties are encoded in the scientific term. This is not a neutral *concept*. The other is being a sensation that is attended to by Mary. This is a neutral *concept*; if the identity is true, it is the neutral *concept* of a physical property. (2001: 205; italics added)

The properties of being B₅₂, and being a sensation that is attended to by Mary are said by Perry to be properties but also concepts. The properties are modes of presentation in the metaphysical sense, but concepts are naturally taken to be or to involve modes of presentation in the cognitive sense. The view he actually argues for is this: “We need instead the topic-neutrality of demonstrative/recognitional concepts” (2001: 205).
end p.262

When I described the metaphenomenal move in the dialectic concerning the Knowledge Argument, I said the phenomenal concept that Mary acquires itself contains or else requires unreducible phenomenality. Why “contains or else requires”? In terms of the CMoP/MMoP distinction: if the CMoP that Mary acquires is partly constituted by an unreducible phenomenal element, then we could say that the concept contains unreducible phenomenality. If the MMoP that is paired with the CMoP involves unreducible phenomenality, one could say that the concept that Mary acquires *requires* an unreducible phenomenal property, as a property of the referent.

In the next section, I will state a version of the Property Dualism Argument in terms of MMoPs. But as we shall see, that argument fails because of what amounts to equivocation: one premise is plausible only if modes of presentation are MMoPs, whereas the other premise is plausible only if modes of presentation are CMoPs. A second version of the Property Dualism Argument will also be couched initially in terms of MMoPs, but that treatment is tactical, and the argument will entail some separate discussion of CMoPs and MMoPs.

I will pause briefly to say where I stand on the main issue. The Property Dualism Argument is concerned with a mind-body identity that says that phenomenal property Q = brain property B₅₂. The worry is that the mode of presentation of Q brings in a nonphysical property. But mode of presentation in which sense? Start with the CMoP. Well, a phenomenal CMoP has a constituent that is phenomenal and is used to pick out something phenomenal. Let me explain.

If I think about the phenomenal feel of my pain *while I am having it*, I can do that in a number of different ways. I could think about it using the description “the phenomenal feel of this pain.” Or I could think about it using the phenomenal feel of the occurring pain itself as part of the concept. But if a token phenomenal feel does double duty in this way (as a token of an aspect of both the pain and our way of thinking of the pain), no extra specter of dualism arises. If the phenomenal feel is a physical property, then it is a physical property even when it (or a token of it) does double duty. The double duty is not required by a phenomenal concept. One could in principle use one phenomenal feel in a CMoP to pick out a different phenomenal feel. For example, the phenomenal feel of seeing green could be used to pick out the phenomenal feel of seeing red if the concept

involves the description “complementary” in the appropriate way. But there is no reason to think that such a use brings in any new specter of dualism.

Move now to the MMoP. We can think about a color in various ways by attending to different properties of that color. I might think of a color via its property of being my favorite color or the only color I know of whose name starts with “r.” Or, I may think about it via its phenomenal feel. And what holds of thinking about a color holds for thinking about the phenomenal feel itself. I can think of it as my favorite phenomenal feel, or I can think about it phenomenally (for example, while looking at the color or imagining it). If the referent is a phenomenal property P, the MMoP might be taken to be the property of being (identical to) P. If P is physical, so is being P. So the MMoP sense generates no new issue of dualism. That is where I stand. The property dualist, by contrast, thinks that there are essential features of

end p.263

modes of presentation that preclude the line of thought that I expressed. That is what the argument is really about.^{13, 14}

I have not given a detailed proposal for the nature of a phenomenal CMoP because my case does not depend on these details. But for concreteness, it might help to have an example. We could take the form of a phenomenal CMoP to be “the experience:____,” where the blank is filled by a phenomenal property, making it explicit how a CMoP might mix descriptive and nondescriptive elements.¹⁵ If the property that fills the blank is phenomenal property P, the MMoP that is paired with this CMoP might be the property of *being* P, and the referent might be P itself.

I will turn now to a bit more discussion of the CMoP/MMoP distinction and then move to stating and refuting the Property Dualism Argument.

Different versions of the Property Dualism Argument presuppose notions of CMoP and MMoP geared to different purposes. I have mentioned two purposes, fixing reference and accounting for cognitive significance. A third purpose—or rather a constraint on a purpose—is the idea that the MMoP is a priori accessible on the basis of the CMoP. And because one cannot assume that these three functions (cognitive significance, fixing reference, a priori accessibility) go together, one wonders how many different notions of CMoP and MMoP there are. Burge (1977) and Byrne and Pryor (2006) give arguments that, although put in different terms, can be used to make it plausible that these three *raison d'être* of modes of presentation do not generally go together. However, I will rebut the Property Dualism Argument without relying—except at one point—on any general claim that this or that function does not coincide with a different function. All of the versions of the CMoP that I will be considering share a notion of a CMoP as a cognitive entity, for example a mental representation. The MMoP, by contrast, is always a property of the referent. One way in which the different *raison d'être* matter is that for fixing reference, the MMoP must not only apply to the referent but uniquely pick it out. Further, it must have been in effect given a special authority in picking out the referent by the subject. But when it comes to cognitive significance, the MMoP need not even apply to the referent (as Byrne and Pryor note in somewhat different terms), so long as it seems to the subject to apply. However, I will not be making use of this difference.

end p.264

Because physicalists say that everything is physical, they are committed to the claim that everything cognitive, linguistic, and semantic is physical. However, not all issues for physicalism can be discussed at once, and since this chapter's focus is on the difficulty that phenomenality poses for physicalism, I propose to assume that the cognitive, linguistic, and semantic features of CMoPs do not pose a problem for physicalism so long as they do not involve anything phenomenal.

I will argue that the key step in the Property Dualism Argument can be justified in a number of ways, assuming rather different ideas of what MMoPs and CMoPs are (so there is really a family of property dualism arguments). There are many interesting and controversial issues about how to choose from various ways of fleshing out notions of CMoP and MMoP. My strategy will be to try to avoid these interesting and controversial issues, sticking with the bare minimum needed to state and critique the Property Dualism Argument. In particular, I will confine the discussion to CMoPs and MMoPs of singular terms, since the mind-body identities I will be concerned with are all of the form of an “=” flanked by singular terms (usually denoting properties). I will not discuss belief contexts or other oblique contexts. The reader may wonder if all these different and underspecified notions of mode of presentation are really essential to any important argument. My view, which I hope this chapter vindicates, is that there is an interesting family of arguments for dualism involving a family of notions of mode of presentation and that this family of arguments is worth spelling out and rebutting.

Am I assuming the falsity of a Millian view, according to which only the referent contributes to what is expressed and modes of presentation do not figure in concepts (thinking of concepts as components of what is expressed)? (See Kripke 1980, p. 20.) Without modes of presentation, the Property Dualism Argument does not get off the ground, so if Millianism assumes that there are no modes of presentation involved in concepts, then I am assuming Millianism is false. However, the view of phenomenal concepts that I will be using has some affinities with a Millian view. In addition, I will consider in the next section a version of the Property Dualism Argument in which metaphysical modes of presentation on both sides of the identity are assumed to be identical to the referent.

Modal arguments for dualism such as Kripke's and Chalmers's attempt to move from epistemic premises to metaphysical conclusions. (For example, the epistemic possibility of zombies is appealed to in order to justify a claimed metaphysical possibility of zombies.) A similar dynamic occurs with respect to the Property Dualism Argument. One way it becomes concrete in this context is via the issue of whether in an identity statement with different CMoPs there must be different MMoPs. That is, is the following principle true?

D(CMoP) → D(MMoP): A difference in CMoPs in the two terms of an identity statement entails a difference in MMoPs.

Prima facie, it seems that the D(CMoP) → D(MMoP) principle is false. Consider the identity “the wet thing in the corner = the thing in the corner covered or soaked with H₂O.” Suppose the CMoP associated with the left-hand side of the identity statement to be the description “the wet thing in the corner.” Take the

end p.265

corresponding MMoP to be the property of being the wet thing in the corner. Analogously for the right-hand side. But the property of being the wet thing in the corner = the property of being the thing in the corner covered or soaked with H_2O . $MMoP_1 = MMoP_2$. That is, there is only one MMoP, even though there are two CMoPs.

Of course, a theorist who wishes to preserve the $D(CMoP) \rightarrow D(MMoP)$ principle, seeing MMoPs as shadows of CMoPs, can postulate different, more finely grained quasi-linguistic-cognitive MMoPs that are individuated according to the CMoPs. There is no matter of fact here but only different notions of CMoP and MMoP geared to different purposes. In the discussion to follow, I will focus on the cognitive significance purpose of the CMoP/MMoP pair, since I think that rationale is the most favorable to the view I am arguing against, that we must—that we are forced to—individuate MMoPs according to CMoPs.¹⁶

Consider the familiar “Paderewski” example. Our subject starts out under the false impression that there were two Paderewskis at the turn of the twentieth century, a Polish politician and a Polish composer. Later, he forgets where he learned the two homographic names and remembers nothing about one Paderewski that distinguishes him from the other. That is, he remembers only that both were famous Polish figures at the turn of the twentieth century. Prima facie, the cognitive properties of the two uses of “Paderewski” are the same. For the referent is the same and every property associated by the subject with these terms is the same. However, there is a cognitive difference. We could give a name to the relevant cognitive difference by saying that the subject has two “mental files” corresponding to the two uses of “Paderewski.” We could regard the difference in mental files as a semantic difference, or we could suppose that semantically the two uses of “Paderewski” are the same, but that there is a need for something more than semantics—something cognitive but nonsemantic—in individuating CMoPs. In either case, there are two CMoPs but only one MMoP, the MMoP being, say, the property of being a famous turn-of-the-twentieth-century Pole named “Paderewski.” Thus “Paderewski = Paderewski” could be informative to this subject, despite identical MMoPs for the two terms.

As Loar (1988) notes, Paderewski-type situations can arise for general terms even in situations in which the subject associates the same description with the two uses of the general term. An English speaker learns the term “chat” from a monolingual French speaker who exhibits cats, and then is taught the term “chat” again by the same forgetful teacher exhibiting the same cats. The student tacitly supposes that there are two senses of “chat” which refer to creatures that are different in some respect that the student has not noticed or perhaps in some respect that the student could not have noticed, something biological beneath the surface that is not revealed in the way they look and act. We can imagine that the student retains two

end p.266

separate mental files for “chat.” Each file has some way of specifying some observable properties of chats, for example that they are furry, purr, are aloof, are called “chat.” Most important, each of the files says that there are two kinds of creatures called “chat”: chats in the current sense are not the same as chats in the other sense. So if the student learns “this chat = this chat,” where the first “chat” is linked to one file, and the second is linked to the other, that will be informative. It is certainly plausible that there are different CMoPs, given that there are two mental files. But the MMoP associated with both CMoPs would seem to be the same—being furry, purring, being aloof, and being called “chat.”¹⁷

It may be objected that there cannot be only one MMoP because explaining cognitive significance requires postulating a difference somewhere; if the difference doesn't lie in the MMoP of the referent, perhaps there are two different MMoPs of that MMoP, or two different MMoPs of the MMoP of the MMoP of the referent.¹⁸ But these higher order MMoPs need not exist! The MMoP of chats in both senses of “chat” is something like being one of two kinds of furry, purring, aloof pets with a certain look. There will not be any further MMoP of that MMoP unless the subject happens to have a thought about the first MMoP. What, then, explains the difference in cognitive significance between the two “chats”? Answer: The difference in the CMoPs, the difference I have given a name to with the locution of different mental files. Objection: “But that difference in CMoP must correspond to a difference in MMoP!” To argue this way is simply to beg the question against the idea that there can be two CMoPs but only one MMoP.

Objection: “But the cognitive difference between the two CMoPs has to correspond to a difference in the world in order to be explanatory. For example, the subject will think, ‘The chat on my left is of a different kind from the chat on my right’.” Answer: No, the example has been framed to rule out this kind of difference. The subject does not remember *any* differences between the two kinds of chats, not even differences in the situations in which he learned the terms.

It may seem that wherever there is a difference in CMoP, there *has* to be *some* difference in MMoP of some kind, for otherwise how would the difference in CMoP ever arise? Thus, corresponding to the different CMoPs “covered with water” and “covered with H₂O,” one might imagine that “water” is learned or applied on the basis of properties such as, for example, being a colorless, odorless, tasteless liquid coming out of the tap, and “H₂O” is learned and applied on the basis of something learned in a chemistry class having to do with hydrogen and oxygen. Similarly, one might say that in the “chat” case, there must be some difference between the property
end p.267

instantiated in the first and second introductions of the word “chat” to the student. For example, perhaps the first one was introduced on a cloudy day and the second on a sunny day. Or at any rate, they were introduced at different times, and so there is a difference in *temporal* MMoPs. For if there were no difference at all in the world, what would explain—that is, explain as rational—why the subject thinks there are different referents? But this reasoning is mistaken. Maybe there has to be some difference in properties in the world that explain the *arising* of the different CMoPs, but that difference can fade away, *leaving no psychological trace*. After the student learns the word “chat” twice, and tacitly assumes that it applies to different animals, the student may forget all the specific facts

concerning the occasions of the learning of the two words, while still tacitly supposing that things that fit “chat” in one sense do not fit it in the other. The *ongoing* use of two cognitive representations corresponding to the two uses of “chat” do not require any *ongoing* difference in MMoPs to be completely legitimate and rational. Likewise for the “Paderewski” example. To suppose otherwise is to confuse ontogeny with metaphysics. The following reply would fit the view of many dualists such as Chalmers and White: But doesn't there have to be a possible world, different from the actual world, that the subject rationally supposes he is in, in which the two CMoPs are CMoPs of different referents? For the subject who believes there are two different Paderewskis, a musician and a politician, the rationalizing world is a world that contains two persons named “Paderewski,” both born around the turn of the century, one famous as a politician, the other famous as a musician. Now in your version of the chat and Paderewski stories as you tell them, you have eliminated all differences in specific properties available to the subject. You have postulated that the subject does not believe that one is a politician and the other is a musician—but the same strategy can be followed all the same. The world that rationalizes the subject's view that there are two Paderewskis is a world in which there are two persons named “Paderewski,” both Europeans born around the turn of the century. The subject knows that there are bound to be many properties that distinguish them (if only their spatial locations), and he can single out two properties in his imagination, X and Y, such that one has property X but lacks Y, the other has property Y but lacks X. If the subject were rationalizing his belief, he could appeal to X and Y, so they can constitute his different MMoPs. One of his MMoPs, call it MMoP_A is X; the other, MMoP_B = Y. The fact that the subject does not know what X and Y are does not change the fundamental strategy of rationalizing the subject's error in which the cognitive difference, CMoP_A vs. CMoP_B, requires a metaphysical difference, that between MMoP_A and MMoP_B.

This territory will be familiar to those who have thought about modal arguments for dualism. The dualist supposes that the conceivability of zombies justifies the claim that there is a possible world in which there is a zombie, and that leads by a familiar route to dualism.¹⁹ The physicalist resists the argument from epistemology to metaphysics in end p.268

that case, and the physicalist should resist it here as well. We can explain the erroneous view that Paderewski is distinct from Paderewski by reference to *epistemic possibilities only*: The epistemically possible situation (not a genuine metaphysically possible world) in which, as one might say, Paderewski is not Paderewski. This is an epistemic situation in which Paderewski—who has property X but not Y (and, as we the theorists might say, is identical to the actual Paderewski)—is distinct from Paderewski, who has property Y but not X (and who, as we the theorists might say, is also identical to the actual Paderewski). Of course there is no such world, but this coherently describable epistemic situation accurately reflects the subject's epistemic state. We need only this coherently describable epistemic situation, not a genuine difference in properties in a genuinely possible world. (I follow the common convention of calling a genuinely possible situation a world and reserving “situation” for something that may or may not be possible.) Likewise for the chat example. *The rationality of error can be explained epistemically*

with no need for metaphysics. This is a basic premise of this chapter, and it links the physicalist position on the Property Dualism Arguments to the physicalist position with regard to the Kripke-Chalmers modal arguments.

Given this principle, I believe that the Property Dualism Argument, the Knowledge Argument, and the familiar modal arguments can be defanged, so the residual issue—not discussed here—is whether this principle is right. Chalmers and White argue that genuine worlds are needed to rationalize the subject's behavior, but I have not seen anything in which they argue against situations as rationalizers.

In my view, the issue I have been discussing is the key issue concerning all forms of the Property Dualism Argument (and some modal arguments for dualism as well). If the $D(\text{CMoP}) \rightarrow D(\text{MMoP})$ principle does not come up in some form or other, the main issue has been skipped.

There is one reason for the view that a difference in CMoPs entails a difference in MMoPs that I have not yet mentioned and will not go into in detail until the “thin/thick” section at the end of the chapter: the view that MMoPs must be thin in the sense of having no hidden essence in order to account for their role in determining reference and explaining cognitive significance.

Of course, as before, those who prefer to see MMoPs as shadows of CMoPs can think of the property of being a chat (relative to the link to one mental file) as distinct from the property of being a chat (relative to the link to the other mental file). That is, the MMoP would be individuated according to the corresponding CMoP to preserve one-to-one correspondence. According to me, one can individuate MMoPs as shadows of CMoPs—or not—but as we will see, the Property Dualist has to insist on individuating MMoPs as shadows of CMoPs.

What about the converse of the cases we have been talking about—one CMoP, two MMoPs? People often use one mental representation very differently in different circumstances without having any awareness of the difference. Aristotle famously used the Greek word we translate with “velocity” ambiguously, to denote both instantaneous velocity and, in other circumstances, average velocity. He did not appear to see the difference. And the Florentine “Experimenters” of the seventeenth century used a term translated as “degree of heat” ambiguously to denote heat and a very different magnitude, temperature. Some of their measuring procedures for detecting “degree of heat” measured heat, and some measured temperature (Block and Dworkin
end p.269

1974). For example, one test of the magnitude of “degree of heat” was whether a given object would melt paraffin. This test measured whether the temperature was above the melting point of paraffin. Another test was the amount of ice an object would melt. This measured amount of heat (Wiser and Carey 1983). One could treat these cases as one CMoP which refers via different MMoPs, depending on context. Alternatively, one could treat the difference in context determining the difference in CMoP, preserving the one-to-one correspondence. This strategy would postulate a CMoP difference that was *not available from the first-person point of view*, imposed on the basis of a difference in the world. That is, it would take a conceptual revolution for theorists of heat phenomena to see a significant difference between their two uses of “degree of heat,” so the cognitive

difference was not one that they could be aware of, given their conceptual scheme. A CMoP difference that is not available to the subject is not acceptable for purposes that emphasize the relevance of the CMoP to the first person.

In what follows, I will assume the existence of independently individuated CMoPs and MMoPs. However, at one crucial point in the dialectic, I will examine whether individuating MMoPs according to CMoPs makes any difference to the argument, concluding that it does not. Why does it matter whether or not there is a one-to-one correspondence between CMoPs and MMoPs? I will now turn to a member of the family of property dualism arguments that turns on this issue. The argument of the next section, or something much like it, has been termed “the property dualism argument” by McGinn (2001), though I think a somewhat different argument is more closely related to what Smart, Perry, and White have in mind, what I will call the “orthodox” property dualism argument. The two arguments depend on nearly the same issues.

E → 2M Version of the Property Dualism Argument

Saul Kripke (1972) argued for dualism as follows. Identities, if true, are necessarily true. But cases of mind without brain and brain without mind are possible, so mind-brain identity is not necessary, and therefore is not true.²⁰ A standard physicalist response is that the mind-body relation is necessary, but appears, misleadingly, to be contingent: there is an “illusion of contingency.” Most of the discussion of an illusion of contingency has focused on the mental side of the identity statement, but Richard Boyd (1980) noted that one way for a physicalist to explain the illusion of contingency of “Q = cortico-thalamic oscillation” would be to exploit the gap between cortico-thalamic oscillation and its mode of presentation. When we appear to be conceiving of Q without the appropriate cortico-thalamic oscillation (e.g., a disembodied mind or a version of spectrum inversion), all we are managing to conceive is Q in a situation in which we are misled by our mode of epistemic access to cortico-thalamic oscillation. What we are implicitly conceiving, perhaps, is a situation in which our functional magnetic resonance scanner is broken. So the physicalist is free to insist that cortico-thalamic oscillation is part of
end p.270

what one conceives in conceiving of Q, albeit not explicitly, and, conversely, Q is part of what one conceives in conceiving of cortico-thalamic oscillation.

But the sole reason for believing in *implicit* commitment to epistemic failure, such as failing brain measurement devices in these thought experiments, is that it avoids the nonphysicalist conclusion, and that is not a very good reason. The conceivability of zombies, inverted spectra, disembodied minds, and so on, does not seem *on the surface* to depend on implicit conceptions of malfunctioning apparatus. For example, it would seem that one could conceive of the brain and its cortico-thalamic oscillation “neat” (as in whiskey served without ice or water)—that is, without conceiving of any particular apparatus for measuring cortico-thalamic oscillation.

However, the idea that one can conceive of cortico-thalamic oscillation “neat” is useful not just in combating Boyd's objection to Kripke's argument for dualism but also in a distinct positive argument for dualism.²¹

Consider an empirical mind-body property identity claim in which *both* terms of the identity—not just the mental term—have MMoPs that are identical to the referent. (MMoPs are, of course, properties, and we are thinking of the referents of mind-body identity claims as properties as well.) McGinn (2001) claims, albeit in other terms, that this would be true for a standard physicalist mind-body identity claim. “It is quite clear that the way of thinking of C-fiber firing that is associated with ‘C-fiber firing’ is simply that of having the property of C-fiber firing,” he writes. “It connotes what it denotes” (294). Is cortico-thalamic oscillation or potassium ion flow across a membrane its own metaphysical mode of presentation? That depends on what a metaphysical mode of presentation is supposed to be, and that depends on the purpose we have for them. I have mentioned three different conceptions of MMoPs, (1) explaining cognitive significance, (2) determining the referent and (3) a priori graspability (on the basis of understanding the term it is the MMoP of).

Suppose we took explaining cognitive significance as primary. How can we explain why “cortico-thalamic oscillation = cortico-thalamic oscillation” is less informative than “Q = cortico-thalamic oscillation”? Do we need to appeal to an MMoP of being cortico-thalamic oscillation for “cortico-thalamic oscillation”? First, if the identity is true, it is not clear that an MMoP of being cortico-thalamic oscillation is of any use. For if the MMoP of Q is *being* Q, then the MMoP of the left-hand side would be the same as for the right-hand side for both the trivial and the cognitively significant identity. Moreover, other MMoPs can explain the difference in cognitive significance. For example, a scientist might conceive of Q from the first-person point of view but think of cortico-thalamic oscillation in terms of the machinery required to detect it. A scientist might even think of it perceptually, in terms of the experience in the observer engendered by the apparatus, as radiologists often say they do in the case of CAT scans.

end p.271

Suppose instead that we take the special reference-fixing authority as the *raison d'être* of the MMoP. This conception has the advantage that if we have given the special reference-fixing authority to an MMoP, then it is a priori graspable that the referent, if it exists, has that property (Byrne and Pryor 2006). Again, it is not very plausible that the MMoP of “cortico-thalamic oscillation” or “potassium ion flow” is being cortico-thalamic oscillation or being potassium ion flow. What would be the point of giving the special reference-fixing authority for “cortico-thalamic oscillation” to the property of being cortico-thalamic oscillation? (Recall that uniquely determining the referent is not enough for reference fixing—the subject has to also have decided [even if implicitly] that that uniqueness property governs the term, as noted by Byrne and Pryor.)

But there is a kind of mind-body identity in which the right-hand term does more plausibly have an MMoP in both the cognitive significance and the reference-fixing sense that is identical to the referent (or at any rate has the relation of *being* X to X), namely a mental-functional identity claim. I will skip the cognitive significance rationale, focusing on determination of reference. What is our way of fixing reference to the property of

being caused by A and B and causing C and D if not that property itself (or the property of having that property itself): that is, being caused by A and B and causing C and D? For many complex functional properties, it is hard to imagine any other reference-fixing property that could be taken very seriously, since it is hard to see how such functional properties could be singled out without singling out each of the causal relations. Further, the functional property would be plausibly a priori graspable on the basis of a typical concept of it. These considerations suggest that a mental-functional identity claim is a better candidate for the kind of identity claim being discussed here than is the standard mental-physical identity claim.

Since the candidate identity claim has to be plausibly empirical, let us think of the physical side as a *psychofunctional* property (see Block 1978, which introduced this term), that is, a functional property that embeds detailed empirical information that can be discovered only empirically. For example, we can take the functional definition to include the Weber-Fechner Law (which dictates a logarithmic relation between stimulus intensity and perceptual intensity). To remind us that we are taking the right-hand side of the identity to be a psychofunctional property, let us represent it as “PF.”

Let our sample mind-body identity be “ $Q = PF$,” where as before, “Q” denotes a phenomenal property. As before, let us use “M” for the metaphysical mode of presentation of Q, and let us assume that $M = \text{being } Q$. Ex hypothesi, the metaphysical mode of presentation of PF is being PF. But since $M = \text{being } Q$, and the MMoP of $PF = \text{being } PF$, if the identity is true ($Q = PF$), it follows that the MMoPs of both sides are the same. (See fig. 12.1.) But if the MMoPs of both sides are the same, then (supposedly) the identity cannot be a posteriori. Here I assume the principle that an empirical identity must have distinct MMoPs for the two sides of the identity. Call that Empirical $\rightarrow$ 2MMoP, or $E \rightarrow 2M$ for short. That would show that the original a posteriori identity claim—which embeds, you will recall, the Weber-Fechner Law and so cannot be supposed to be a priori—cannot be true. Psychofunctionalism is thus refuted (or so it may seem).

end p.272

**Metaphysical
mode of
presentation
of Q**

||
Q

■

**Metaphysical
mode of
presentation
of PF**

||
PF

Figure 12.1: Empirical? 2MMoP Argument for DualismMMoP (i.e. metaphysical mode of presentation) of $Q = \text{being } Q$, MMoP of $PF = \text{being } PF$, so if it is true that $Q = PF$, then the MMoP of $PF = \text{the MMoP of } Q$. But if the two MMoPs are the same, the identity is supposed to be a priori. However, since the identity is not a priori, the argument concludes, it is not true. The vertical '=' signs represent the relation between X and being X .

The upshot would be that if we want a functionalist mind-body identity thesis, it can only be a priori (in which case deflationism—in the sense of conceptual reductionism about consciousness—holds). Or if we reject deflationism, the upshot is that functionalist mind-body identity is false (i.e., the relevant form of dualism is true). So the conclusion is the same as that of the Property Dualism Argument, but restricted to functionalist mind-body identity claims: only dualism and deflationism are viable.

Why accept the $E \rightarrow 2M$ Principle? Suppose that different CMoPs for the two terms of the identity entail different MMoPs (i.e., the $D(\text{CMoP}) \rightarrow D(\text{MMoP})$ principle). An empirical identity requires different CMoPs, since, it may be said (but see below), if two of one's terms have the same cognitive significance, that fact is a priori available to the subject. An empirical identity requires different CMoPs, different CMoPs require different MMoPs, so an empirical identity requires different MMoPs. It would follow that an empirical identity requires different MMoPs. This is one way of seeing why the considerations of the last section about the one-to-one correspondence between CMoPs and MMoPs matter for dualism.²²

You will not be surprised to learn that my objection to the argument is to the $E \rightarrow 2M$ Principle and the claim that different CMoPs require different MMoPs
end p.273

that engenders the $E \rightarrow 2M$ Principle. As I mentioned, a priority is better taken to be a matter of sameness of CMoPs, not a matter of sameness of MMoPs. In the example given above, before the subject learns that there is only one kind of creature called “chat,” he has two CMoPs but only one MMoP.²³

The argument could be resuscitated if the CMoP of each side were identical to the referent. But at least on the right-hand side, this seems like a category mistake: our concept of a psychofunctional state (or something cognitive about it) is a poor candidate for identity with the psychofunctional state itself.

In comments on this chapter, David Chalmers suggested a variant of the $E \rightarrow 2M$ Argument. Instead of “ $Q = PF$,” consider “ $Q = P$,” where P is a physical property. Assume the $E \rightarrow 2M$ Principle—that an empirical identity must have distinct MMoPs for the two sides of the identity. If “ $Q = P$ ” is empirical, then it follows that the MMoP of Q is distinct from any MMoP of a physical property. For if the MMoP of P is just P and the MMoP of Q is just Q , and since the $E \rightarrow 2M$ principle requires that the two MMoPs be distinct, it follows by transitivity of identity that Q must be distinct from P , and so dualism is true.

My objections to this variant are, as before:

1. The argument assumes the $E \rightarrow 2M$ principle in the first step, in which it is argued that the MMoP of Q is distinct from any MMoP of any physical property, and as mentioned above, I reject the $E \rightarrow 2M$ principle.
2. The argument presupposes the view that it is reasonable to take the MMoP of a physical property, P, to be just P itself. (It would be better to take it to be *being P*, but I will ignore this glitch.) As I emphasized above, I find this doubtful for physical properties, although more plausible for functional properties. So at most, the argument is an argument against empirical functionalism (psychofunctionalism) rather than against physicalism.

Back to Stating the Orthodox Property Dualism Argument

The $E \rightarrow 2M$ argument raises many of the same issues as, but is not quite the same as, the argument that Smart, Perry, and White are talking about.

To frame the orthodox Property Dualism Argument, we need to use a contrast between deflationism and phenomenal realism about consciousness.²⁴ In its strong end p.274

form, deflationism is *conceptual reductionism* concerning concepts of consciousness. More generally, deflationism says that a priori or at least armchair analyses of consciousness (or at least armchair sufficient conditions) can be given in nonphenomenal terms, most prominently in terms of representation, thought, or function.²⁵ (If the analyses are physicalistic, then deflationism is a form of what Chalmers (1996) calls type-A physicalism.) The deflationist says phenomenal properties and states do exist, but that commitment is “deflated” by an armchair analysis that reduces the commitment. The conclusion of the orthodox Property Dualism Argument is that physicalism and phenomenal realism are incompatible: the phenomenal realist must be a dualist, and the physicalist must be a deflationist.

In what follows, I will drop the term “orthodox” and refer to the argument I am spelling out simply as the Property Dualism Argument.

The Property Dualism Argument in the form in which I will elaborate it depends on listing all the leading candidates for the nature of the MMoP of the mental side. My emphasis on the MMoP at the expense of the CMoP is artificial but has some dialectical advantages. The metaphenomenal move is what is really being explored, the view that with the statement of mind-body identity, either or both of the MMoP or the CMoP brings in unreducible phenomenality. Most of the issues that come up with respect to the MMoP could also have been discussed with respect to the CMoP. In rebutting the Property Dualism Argument, I will go back to the CMoP occasionally.

Recall that the phenomenal side (which I will always put on the left side of the sentence on the page) of the identity is Q. Let the metaphysical mode of presentation of Q be M (for *mental*, *metaphysical* and *mode of presentation*). The basic idea of the Property Dualism Argument is that even if Q is physical, there is a problem about the physicality

of M. I will discuss five proposals for the nature of M. M might be (one or more of) the following:

1. mental
2. physical
3. nonphysical
4. topic-neutral, or
5. nonexistent (i.e., the reference is “direct” in one sense of the term)

Here is a brief summary of the form of the argument. Proposal 1 is correct, but it's useless because both the physicalist and the dualist will agree on it. The problem for the physicalist is to show how M can be both physical and mental. Proposal 2 is (supposed to be) ruled out by the arguments given below, which will be the main topic of the rest of this chapter. Proposal 5 changes the subject by stipulating a version of the original property identity “Q = cortico-thalamic oscillation” in which Q is not picked out by a genuine phenomenal concept. So the remaining options are the dualist option (3), and the topic-neutral option (4). White (1986) argues that option 4 is deflationist, reasoning as follows: The topic-neutral
end p.275

properties that are relevant to the mind-body problem are functional properties. If M, the metaphysical mode of presentation of Q, is a topic-neutral and therefore (according to White) functional property, then that could only be because the phenomenal concept has an a priori functional analysis. For example, the concept of pain might be the concept of a state that is caused by tissue damage and that causes certain reactions, including interactions with other mental states. But an a priori functional analysis is deflationist by definition. The upshot is supposed to be that only proposals 3 and 4 remain; 3 is dualist and 4 is deflationist. The conclusion of the Property Dualism Argument is that we must choose between dualism and deflationism: phenomenal realist physicalism is not tenable. Of course, the argument as I have presented it makes the title “Property Dualism Argument” look misguided. Anyone who does take the argument to argue for dualism would presumably want to add an argument against deflationism. However, Smart and Armstrong (1968) (and in a more convoluted version, David Lewis (1980)) used the argument the other way around: the threat of dualism was brought in to argue for deflationism. Their view is that “pain” contingently picks out a physical state, for “pain” is a nonrigid designator whose sense is *the item with such and such functional role*. But the view that stands behind this picture is that the nature of the mental is given a priori as functional. “Pain” is a nonrigid designator, but what it is to have pain, that which cases of pain all share in virtue of which they are pains, is a certain functional property, and that functional property can be rigidly designated by, for example, the phrase “having pain.”

²⁶ So the view is a version of deflationism.

White (1986) added an antidualist premise to the argument whose conclusion is *dualism or deflationism*, but in White (2006) and his chapter for this book (chap. 11), he drops that premise, arguing instead for dualism. My own point of view, the view I'm arguing for in this chapter, is phenomenal realist and physicalist, the very combination that the

Property Dualism Argument purports to rule out. (Though see Block 2002 for a different kind of doubt about this combination.) As we will see when I get to the critique of the Property Dualism Argument, the argument fares better as an argument for dualism than for deflationism, so the name of the argument is appropriate.

There are some well-known problems concerning the notion of a physical property.²⁷ But not all philosophy concerned with physicalism can be about the problem
end p.276

of how to formulate physicalism. For some purposes, physicalism is clear enough.²⁸ In particular, the debate about the Property Dualism Argument seems relatively insensitive to issues about what exactly physicalism comes to. (If not, that is an objection to what follows.)

I will take the notions of physicalistic vocabulary and mentalistic vocabulary to be unproblematic. A physical property is a property canonically expressible in physicalistic vocabulary. (I won't try to explain "canonically.") For example, the property of being water is a physical property because that property = the property of being H₂O. The predicate "___ is H₂O" is a predicate of physics (or anyway physical science), the property of being H₂O is expressed by that predicate, and so is the property of being water, since they are the same property. (Note that the relation of "expression" is distinct from referring.) A mentalistic property is a property canonically expressible in mentalistic vocabulary. "___ is a pain" is a mentalistic predicate and thus expresses (or connotes) a mental property (that of being a pain). A nonphysical property is a property that is not canonically expressible in physicalistic vocabulary. (So physicalism dictates that mental properties are canonically expressible in both physicalistic and mentalistic vocabularies.) I don't know if these notions can ultimately be spelled out in a satisfactory manner, but this is another of the cluster of issues involved in defining physicalism that not every work concerning physicalism can be about.

Smart said that a topic-neutral analysis of a property term entails neither that the property is physical nor that it is nonphysical. It would not do to say that a
end p.277

topic-neutral property is expressible in neither physicalistic nor nonphysicalistic terms, because if physicalistic terms and nonphysicalistic terms are all the terms there are, there will be no such properties. The key kind of topic-neutral property for present purposes is a functional property, a second-order property that consists in the having of certain other properties that are related to one another (causally and otherwise) and to inputs and outputs, all specified nonmentalistically. One could say that a topic-neutral property is one that is expressible in terms of logic, causation, and non-mentalistically specified input-output language. The question may arise as to whether these terms are to be counted as part of physicalistic vocabulary or not. But for my purposes here, I will leave that issue undecided.

I will briefly sketch each of the proposals mentioned above for the nature of M (the metaphysical mode of presentation of Q, which was introduced in the sample identity "Q = cortico-thalamic oscillation") from the point of view of the Property Dualism

Argument, adding some critical comments at a few places. Then, after a section on phenomenal concepts, I will rebut the Property Dualism Argument.

Proposal 1. M is Mental

If M is mental, then the same issue of physicalism arises for M, the metaphysical mode of presentation of Q, which arises for Q itself. It isn't that this proposal is false, but rather that it presents a challenge to the physicalist of showing how it could be true.

Proposal 2. M is Physical

The heart of the Property Dualism Argument is the claim that M cannot be physical. I will discuss three arguments for that claim.²⁹ The first proceeds as follows. If M is physical, it will not account for cognitive significance; specifically, it will not account for the informativeness of identities and the possibility of rational error. For example, suppose the subject rationally believes that Q is instantiated here and now but that cortico-thalamic oscillation is absent. He experiences Q but also has evidence (misleading evidence, according to the physicalist) that cortico-thalamic oscillation is absent. We can explain rational error by appeal to two different MMoPs of the referent, only one of which is manifest. Let us take the metaphysical mode of presentation of the right-hand side of the mind-body identity "Q = cortico-thalamic oscillation" to be a matter of the instrumentation that detects cortico-thalamic oscillation. We can think of this instrumentation as keyed to the oxygen uptake by neural activity. (Functional magnetic resonance is a form of brain imaging that detects brain activity via sensitivity to metabolism of the oxygen that feeds brain activity.)

The focus of this argument is on the left-hand side, the metaphysical mode of presentation of Q, namely M. According to the argument, if M is physical, it cannot serve the purpose of explaining rational error. For, to explain rational error, we require a metaphysical mode of presentation that makes rational sense of the subject's point of view. But the physical nature of M is not available to the subject. (The subject can be presumed to know nothing of the physical nature of M.) The problem could be solved if there was a mental mode of presentation of M itself,

end p.278

call it "M*." But this is the first step in a regress in which a physical metaphysical mode of presentation is itself presented by a mental metaphysical mode of presentation. For the same issue that arose for M will arise all over again for M*. Explaining rational error requires two modes of presentation, the manifestation of which is available to the first person at some level or other, so postulating a physical metaphysical mode of presentation just takes out an explanatory loan that has to be paid back at the level of modes of presentation of modes of presentation ... , and so on. The upshot is that physical metaphysical modes of presentation do not pass the test imposed by one of the stipulated purposes of metaphysical modes of presentation.

There is also a related non-regress argument: if M is physical, a subject could believe he is experiencing Q, yet not believe he is in a state that has M. But there can be no epistemic gap of this sort between the metaphysical mode of presentation of a phenomenal property and the property itself.

Another argument that M cannot be physical is given by White (1986). He notes, plausibly enough, that

Since there is no physicalistic description that one could plausibly suppose is coreferential a priori with an expression like "Smith's pain at t," no physical property of a pain (i.e., a brain state of type X) could provide the route by which it was picked out by such an expression. (1986: 353; 1997: 706)

Or in the terms of this chapter, there is no physicalistic description that one could plausibly suppose is coreferential a priori with a mentalistic expression such as "Q", so no physical property could provide the route by which it was picked out by such an expression. The property that provides the route by which Q is picked out by "Q" is just the metaphysical mode of presentation (on one way of understanding that term) of Q, that is, M. So the upshot is supposed to be that M cannot be physical because there is no physicalistic description that is coreferential a priori with a phenomenal term.

A third argument that M cannot be physical is that MMoPs must be "thin." We can take a thin property to be one that has no hidden essence. "Thick" properties include Putnamian natural kinds such as water. According to the property dualist, the explanatory purpose of MMoPs precludes thick properties serving as modes of presentation. For, it might be said, it is not *all* of a thick property that explains rational error but only an *aspect* of it. The same conclusion can be reached if one stipulates that the MMoP is a priori available on the basis of the CMoP. Since hidden essences are never available a priori, hidden essences cannot be part of MMoPs. I will indicate later how the claim that MMoPs must be thin can be used to argue against the phenomenal realist physicalist position. This consideration can also be used to bolster the regress argument and the argument of the last paragraph.

I said earlier that the standard reply to Jackson's argument attempts to substitute a dualism of concepts for a
end p.279

dualism of properties and facts. And then I noted that the objection that is exploited by both the Knowledge Argument and the Property Dualism Argument is that the dualism of concepts is held to *require* a dualism of properties and facts. Thin MMoPs are in effect individuated according to the corresponding CMoPs. So the attempt to substitute a dualism of concepts for a dualism of properties and facts is opposed by the claim that properties and facts should be individuated according to concepts, and so if Mary acquires a new concept, she acquires a concept that involves new properties and facts. Earlier I discussed the $D(\text{CMoP}) \rightarrow D(\text{MMoP})$ principle, suggesting that there could be cases of two CMoPs with the same MMoP. One example was the identity "the thing in the corner covered with water = the thing in the corner covered with H_2O ." The CMoP associated with the left-hand side is the description "the thing in the corner covered with water," and the corresponding MMoP is the property of being the thing in the corner covered with water. Analogously for the right-hand side. But the property of being the

thing in the corner covered with water = the property of being the thing in the corner covered with H₂O, so there is only one MMoP. But if MMoPs cannot be “thick,” being covered with water cannot be an MMoP. The relevant MMoP would have to be some sort of stripped down version of being covered with water that does not have a hidden essence.³⁰

These three arguments are the heart of the orthodox Property Dualism Argument. I regard the three arguments as appealing to MMoPs in different senses of the term, and when I critique these three arguments later, I will make that point more explicitly. In my critique, I will argue that two of the arguments do not stand on their own, but rather presuppose the third (“thick/thin”) argument. Then I will examine that argument.

Proposal 3. M is Nonphysical

If M is nonphysical, dualism is true. So this proposal will not preserve the compatibility of phenomenal realism with physicalism and will not be considered further here.

Proposal 4. M is Topic-Neutral

In effect, I covered this topic earlier in my discussion of Perry. A genuinely phenomenal concept is required for getting the Property Dualism Argument (and the Mary argument) off the ground, so a topic-neutral concept will not do.

Proposal 5. There Is No M: The Relation between Q and Its Referent Is “Direct” in One Sense of the Term

A phenomenal concept is a phenomenal way of thinking of a phenomenal property. Phenomenal properties can be thought about using nonphenomenal concepts of them, for example, the concept of the property occurring at 5 p.m. As I've said, the Property Dualism Argument requires a phenomenal concept in my sense of the
end p.280

term, and so if the mind-body identity at issue does not make use of a phenomenal concept, the property dualist will simply substitute a mind-body identity that does make use of a phenomenal concept. Of course, if it could be shown that there could not be any phenomenal concepts, then the Property Dualism Argument will fail. But I believe in phenomenal concepts and so will not discuss this view further.

Phenomenal concepts are often said to refer “directly,” but what this is often taken to mean in philosophy of mind discussions is not that there is no metaphysical mode of presentation, but rather that the metaphysical mode of presentation is a necessary property of the referent.

Loar (1990) says:

Given a normal background of cognitive capacities, certain recognitional or discriminative dispositions suffice for having specific recognitional concepts. ... A recognitional concept may involve the ability to class together, to discriminate, things that have a given objective property. Say that if a recognitional concept is related thus to

a property, the property triggers applications of the concept. Then the property that triggers the concept is the semantic value or reference of the concept; the concept directly refers to the property, unmediated by a higher order reference-fixer.³¹

Consider the view that a phenomenal concept is simply a recognitional concept understood as Loar suggests, whose object is a phenomenal property that is a physical property. I don't know if this would count as a concept that has no metaphysical mode of presentation at all, but certainly it has no phenomenal metaphysical mode of presentation, and so is not a phenomenal concept in the sense required for the Property Dualism Argument. For one can imagine a case of totally unconscious triggering of a concept by a stimulus or by a brain state. As Loar notes, there could be an analog of blindsight in which a self-directed recognitional concept is triggered blankly, without any phenomenal accompaniment. (Of course this *need* not be the case—the brain property doing the triggering could itself be phenomenal, or else the concept triggered could be phenomenal. In either case, phenomenality would have to be involved in the triggering of the concept.) And for this reason, Loar (1990: 98; 1997: 603) argues, a phenomenal concept is not merely a self-directed recognitional concept.

To sum up, the central idea of the Property Dualism Argument (and the Knowledge Argument) is the metaphenomenal move, the idea that in thinking about a phenomenal property, a further phenomenal property must be brought in as part of the CMoP or with the MMoP and that this further phenomenal property poses a special problem for physicalism because of its connection to a mode of presentation. There are three functions of modes of presentation on one or another conception of them that putatively lead to this resistance to physicalism: explaining
end p.281

cognitive significance, determining reference, and providing a priori availability on the basis of understanding the term.

The Property Dualism Argument says that in the identity “Q = cortico-thalamic oscillation,” the metaphysical mode of presentation of Q (viz., M) must be either mental, or physical, or nonphysical, or topic-neutral, or “direct,” in which case there is no metaphysical mode of presentation. The mental proposal is supposed to be useless. The physical proposal is supposed to be ruled out because there is no a priori available physicalistic description of Q, thanks to supposed regress, and because the metaphysical mode must be “thin.” The “direct reference” proposal appears to be ruled out by the fact that the concept of Q needed to get the argument off the ground is a phenomenal concept with a phenomenal metaphysical mode of presentation. So the only proposals for M that are left standing are the nonphysical and topic-neutral proposals. The topic-neutral proposal involves a form of deflationism. So the ultimate metaphysical choice according to the Property Dualism Argument is between deflationism and dualism. The upshot is that the phenomenal realist cannot be a physicalist. The argument is a way of making the metaphenomenal move described earlier concrete: the statement of a mind-body identity claim is supposed to be self-defeating because the MMoP (or the CMoP—but I have focused on the MMoP) of the phenomenal term of the identity is supposed to bring in unreducible phenomenality. The only way to avoid that unreducible phenomenality is to give a deflationist analysis; the alternative is dualism.

Objections Concerning Phenomenal Concepts

The notion of phenomenal concept that I've used is based on the observation that there is a fundamental exercise of it in which a token of a phenomenal property can serve in thought to represent a phenomenal property. In such a case, there is a phenomenal property that is part of the CMoP. There is a special case that I mentioned earlier in which a token of a phenomenal property can serve in thought to represent that very phenomenal property. In such a case, the phenomenal property does double duty: as part of the concept and also as the referent of that concept. Before I go on to rebut the Property Dualism Argument, I will briefly consider two objections to this conception of a phenomenal concept.

Objection (put to me by Kirk Ludwig): I can truly think, "I am not having an experience as of red now," using a phenomenal concept of that experience, but that would not be possible on your view of what phenomenal concepts are.

Reply: Ludwig is right that one can truly think, "I am not having a red experience now," using a phenomenal concept of that experience. As I mentioned, a phenomenal concept has nonfundamental uses in which there is nothing phenomenal going on in the exercise of the concept. But even in one of the fundamental uses in which a token of an experience as of red is being used to represent that experience, it is possible to think a false thought to the effect that one is not having that experience. For example, one might set oneself to think something that is manifestly false, saying to oneself, "I am not having an experience as of red now,"
end p.282

using a phenomenal concept—in my heavy-duty sense of phenomenal concept—of the experience.

Objection: On your view, a phenomenal property does double duty—it is the referent but also is part of the mode of presentation of that referent. But if physicalism is true, cortico-thalamic oscillation would be part of its own mode of presentation. Does that really make sense?

Reply: The claim is not that the right-hand side of the identity "Q = cortico-thalamic oscillation" has an associated mode of presentation (CMoP or MMoP) that involves cortico-thalamic oscillation. I have been supposing that the modes of presentation of the right-hand side have to do with the physical properties of oxygen metabolism that are exploited by scanning technology. Modes of presentation—both cognitive and metaphysical—are modes of presentation associated with *terms* or the concepts associated with the terms, and the identity involves *two* terms. There is no conflict with the indiscernibility of identicals if one keeps use and mention distinct. That is, cortico-

thalamic oscillation is part of its own mode of presentation only as picked out by the phenomenal concept of it.³²
end p.283

Critique of the Property Dualism Argument

The Property Dualism Argument says that the metaphysical mode of presentation of Q, namely M, cannot be physical (using the identity “Q = cortico-thalamic oscillation” as an example). I mentioned three (subsidiary) arguments to that effect, a regress argument, an argument concerning a priori availability, and an argument based on the thin/thick distinction. I also mentioned three different *raison d'être* of modes of presentation, each of which could be used with respect to any of the three arguments, yielding, in principle, nine distinct arguments—even eighteen if one counts the CMoP/MMoP dimension—making refutation potentially unmanageable. I will try to finesse this multiplicity by taking the strongest form of each argument, and bringing in the other *raison d'être* as they are relevant. (I have already mentioned my focus on the MMoP in most of the argument at the expense of the CMoP.) The exposition of the argument has been long, but the critique will be much shorter. As we will see, the first two arguments do not really stand alone, but require the thin/thick argument. My critique of the thin/thick argument is aimed at depriving the conclusion of support rather than outright refutation.

Regress

The first argument mentioned earlier against the physical proposal is a regress argument. The idea is that if M is physical, it cannot account for cognitive significance (informativeness). For example, suppose the subject rationally believes that he has Q but not cortico-thalamic oscillation. As noted earlier, there can be rational error in supposing A is present without B when in fact A = B. That error can be explained if, at a minimum, there is a metaphysical mode of presentation of A, MMoP_A and a metaphysical mode of presentation of B, MMoP_B, such that MMoP_A is manifest, and MMoP_B is not. Applied to the case at hand, the physicalist thesis that Q = cortico-thalamic oscillation, let us assume that the MMoP of “cortico-thalamic oscillation” is the one mentioned earlier having to do with oxygen uptake by neural processes that affects a brain scanner. It is the other metaphysical mode of presentation that is problematic: M, the metaphysical mode of presentation of the left-hand side of the identity. The property dualist says that if M is physical, then M cannot account for cognitive significance because the subject need have no access to that physical description just by virtue of being the subject of that metaphysical mode of presentation. The problem could be solved if there was a *mental* mode of presentation of M *itself*, call it “M*.” But this is the first step in a regress in which a metaphysical mode of presentation that is physical is itself presented by a metaphysical mode of presentation that is mental. For the same issue will arise for M* that arose for M. Accounting for the different cognitive significances of the two sides of

an identity statement requires two modes of presentation that are available to the first person at *some level or other*, so postulating a physical metaphysical mode of presentation just takes out an explanatory loan that has to be paid back
end p.284

at the level of modes of presentation of modes of presentation, and modes of ... and so on.

This argument begs the question. It supposes that if M is physical, it could not serve to account for cognitive significance, since accounting for cognitive significance requires a mental MMoP. But the physicalist thesis is that M is *both* mental and physical, so the physicalist will not be concerned by the argument.³³ Thus, the regress argument in the form I described is like the old objection to physicalism that says that brain states involve the instantiation of electrochemical properties, but since pain does not involve the instantiation of such properties, pain can't be a brain state.

Of course, if MMoPs must be thin, then M, which is an MMoP, cannot have a hidden physical nature, and so it cannot be both mental and physical. But if that is the claim, the regress argument depends on the "thick/thin" argument to be discussed below, and it does not stand on its own.

I assumed that the MMoP of "cortico-thalamic oscillation" is unproblematic, having to do, for example, with oxygen metabolism as a result of brain activity. But the property dualist may say that this MMoP does not uniquely determine the referent and need not be a property to which the subject has given a special reference-fixing authority. (I will use the phrase "fixes the referent" to mean uniquely determines the referent and has been given the special authority.) Why is this a reply to my point concerning the question-begging nature of the regress argument? The question arises: if the regress argument's appeal to cognitive significance requires an MMoP for "cortico-thalamic oscillation" that *does* fix the referent, what would that MMoP be? Someone could argue that that MMoP could only be the property *being cortico-thalamic oscillation itself*. And then it could be claimed that both sides of the identity statement are such that the MMoP of that side is identical with the referent. And this may be said to lead to dualism via the route canvassed earlier in the section on the E → 2M Argument. (If the MMoP of the right-hand side of an identity of the form X = Y is being Y, and the MMoP of the left-hand side is being X, then, if it is true that X = Y, it follows that being X = being Y, so the MMoPs of the two sides are the same. The E → 2M argument goes on to conclude that the identity must therefore be a priori if true, so therefore false.) I will not go into the matter again, except to note that it cannot be assumed that a property of the referent that accounts for cognitive significance also fixes the referent, and what counts in this argument is cognitive significance. As Burge (1977) and Byrne and Pryor (2006) note, it is easy to see that properties of the referent that account for cognitive significance need not fix the referent. As Burge says, the determination of reference depends on all sorts of nonconceptual contextual factors that "go beyond what the thinker 'grasps' in thought" (358). Byrne and Pryor give the example that *being a raspy-voiced singer* may give the cognitive significance for "Bob Dylan," even though there are other raspy-voiced
end p.285

singers. And being a raspy-voiced singer need only be a property that the subject saliently associates with the referent, not a property to which the subject has given the special authority.³⁴ (This, incidentally, is the one point at which I appeal to general considerations about whether the three *raison d'être* for modes of presentation mentioned earlier go together.)

In sum, the regress argument depends on the “thin/thick” argument and does not stand alone.

To avoid confusion, let me just briefly mention something the Property Dualism Argument is *not*. Someone might ask the question, In the identity “A = B,” how does one think of the metaphysical mode of presentation of A, $MMoP_A$? Doesn't one need a metaphysical mode of presentation of $MMoP_A$, which we could call $MMoP_A^*$? And another of that, $MMoP_A^{**}$? And the series won't end without some kind of “direct acquaintance” that does not require an $MMoP$ (Cf. Schiffer 1990: 255). Answer: One does not *need* to think about $MMoP_A$ to use $MMoP_A$ to think about A. However, if one does *happen* to want to think about $MMoP_A$, then one does need a concept of $MMoP_A$ with its own $MMoP$. “And don't we have to have a way of thinking of $MMoPs$ that don't involve further $MMoPs$ to avoid a regress?” Answer: No. To frame a thought about anything, we need a concept of it, including both a $CMoP$ and an $MMoP$. To think about that $CMoP$, we need a further concept of it, and to think about the $MMoP$ we need a further concept of that. Every layer of thinking about a concept of a concept of ... makes it harder to do the mental gymnastics required to form the thought, and for most people, the ability to think these ever more complex thoughts will run out pretty quickly. So there is no regress; the mental gymnastics are voluntary. By contrast, the allegation of the regress argument that is part of the Property Dualism Argument is that we *must* go up a level in order to explain cognitive significance at the preceding level. This is logically required and not just voluntary mental gymnastics.

A Priori Availability

The second argument presented above was that (to quote White 1986), Since there is no physicalistic description that one could plausibly suppose is coreferential a priori with an expression like “Smith's pain at t,” no physical property of a pain (i.e., a brain state of type X) could provide the route by which it was picked out by such an expression. (353)

So the $MMoP$ of the mental side of a mind–body identity claim could not be physical.
end p.286

The first thing to notice about this argument is that if “Smith's pain at t” is taken to be the relevant mental concept in the Property Dualism Argument, it has the flaw of being purely linguistic and not a phenomenal concept of the sort I have argued is required for the argument. Still, it might seem that the argument goes through, for a genuinely

phenomenal concept does not make a physical description of anything that could be called the route of reference any more available a priori than does the description “Smith’s pain at t.”

Note that the *raison d’être* of modes of presentation assumed here is not the cognitive significance appealed to in the regress argument but rather: the property of the referent (i.e., MMoP) that provides “the route by which it is picked out.” What is “the route by which it is picked out”? I think the right thing to mean by this phrase is what I have called fixing the referent, but I doubt that anything hangs on which of a number of candidates is chosen. Consider a case in which the subject conceives of the referent as being the local wet thing. Let us suppose:

- The property of being the local wet thing is a priori available to the subject on the basis of understanding the term and therefore grasping its CMoP.
- The property of being the local wet thing uniquely determines the referent.
- The subject has given this property the special reference-fixing authority mentioned earlier.

My strategy is to concede all that could reasonably be said to be involved in reference fixing and to argue that nonetheless the argument does not work. For being wet = being at least partially covered or soaked with H_2O . But the subject whose metaphysical mode of presentation it is need not have a priori access to “being at least partially covered or soaked with H_2O ” or know a priori that this physical description is coreferential with the original description. The subject can give the property of being the local wet thing the special reference-fixing authority, and thus have that property a priori available from the first-person point of view, without ever having heard of “ H_2O .” I hereby stipulate that the name “Albert” is the name of the local wet thing. In virtue of my grasp of the term “Albert,” the property of Albert’s being the local wet thing is a priori available to me. Also, I have stipulated that the property of being the local wet thing has the special reference-fixing authority. But I can do all that without knowing *all* descriptions of that property. That property *can be and is physical* even though I do not know, and therefore do not have a priori available, its physicalistic description.

Earlier, I considered the idea that MMoPs should be individuated according to CMoPs and thus that the property of being the local wet thing—considered as an MMoP-individuated-according-to-CMoP—is not identical to the property of being covered or soaked with H_2O because the *terms* “water” and “ H_2O ” are not identical. And of course this way of individuating the MMoP would provide an objection to the argument of the last paragraph.

However, the question then arises of what it is for such properties to be physical and what the physicalist’s commitments are with respect to such properties. I believe that this question is best pursued not by inquiring about how to think of such strange entities as MMoPs-individuated-according-to-CMoPs but by focusing on the CMoPs themselves. And a further reason for turning the focus to CMoPs is that

although the subject need have no a priori access to the physical descriptions of the physical properties that provide the metaphysical route of access, it may be thought that

this is not so for CMoPs. After all, CMoPs are certainly good candidates for something to which we have a priori access!

Let us distinguish two things that might be meant by saying that a CMoP (or MMoP) is physical. First, one might have an *ontological* thesis in mind—that the CMoP (or MMoP) is identical to a physical entity or property or some conglomeration involving physical properties or entities. In this sense, a CMoP (or MMoP) can be physical whether or not the subject has a priori access to any physicalistic description of it. (The issue with which the Property Dualism Argument is concerned is whether phenomenal properties are, ontologically speaking, physical properties. I said at the outset that the issue of whether the cognitive apparatus involved in a CMoP is ontologically physical should be put to one side [except to the extent that that apparatus is phenomenal]. My rationale, you will recall, is that although there is an important issue as to whether physicalism can handle cognitive [and semantic] entities or properties, in a discussion of whether *phenomenal* properties are physical, a good strategy is to suppose that nonphenomenal cognitive and semantic entities are not physically problematic.)

A second interpretation of the claim that a CMoP is physical is that it is *explicitly* physical or explicitly analyzable a priori in physical terms. In this chapter, I have been using “physicalistic” to mean explicitly physical. It is not obvious what it would mean to say that an MMoP is or is not physicalistic (since it is not a cognitive, linguistic, or semantic entity), but it does make sense to say that something that involves conceptual or linguistic or semantic apparatus is or is not physicalistic. For example, the CMoP “being covered with water” is not physicalistic (at least if we restrict physics to microphysics), whereas “being covered with H₂O” is physicalistic.

Is the CMoP of a phenomenal concept physical? Physicalistic? Recall that according to me, a phenomenal concept uses a (token of a) phenomenal property to pick out a phenomenal property. Thus the CMoP of a phenomenal concept contains a nondescriptive element: a phenomenal property. And a phenomenal property is certainly not *explicitly* physical (physicalistic), that is, it does not contain conceptual apparatus or vocabulary of physics. A phenomenal property is not a bit of conceptual apparatus, and it contains no conceptual apparatus. So focusing on the “physicalistic” sense of “physical,” the CMoP of a phenomenal concept is not physical. Must the physicalist therefore admit defeat? Hardly, for physicalism is not the doctrine that everything is explicitly physical. Physicalism does not say that all descriptions or conceptual apparatus are couched in physical vocabulary or analyzable a priori in physical vocabulary. Physicalists allow that there are domains of thought other than physics. Physicalists do not say that economics, history, and anthropology use physicalistic vocabulary or conceptual apparatus. This is an absurd form of conceptual or terminological reductionism that cannot be equated with physicalism.

Physicalism does not require that the CMoP of a phenomenal concept be physicalistic, but it does require that it be (ontologically) physical. Is it physical? That depends partly on whether all semantic and cognitive apparatus is physical, an issue that I have put aside for the purposes of this chapter. So the remaining issue is whether the phenomenal property that is part of the CMoP is physical. And that, of
end p.288

course, is the very issue of physicalism versus dualism that is our subject matter. The Property Dualism Argument cannot *assume* that it is not physical; that is what the argument is supposed to show.

Where are we? Here is the dialectic: The property dualist says that in order for physicalism to be true, the physical description of the property that provides the route of reference (of the phenomenal term in a phenomenal-physical identity) has to be a priori available to the subject; it is not a priori available; so physicalism is false. I pointed out that even on very liberal assumptions about the role of the MMoP, a priori availability of a physical description of a physical property is an unreasonable requirement. But then I imagined a property dualist reply which said that I had failed to individuate the MMoP according to the CMoP. I then suggested that we eliminate the middleman, looking at the CMoP itself instead of considering the MMoP-individuated-according-to-the-CMoP. I pointed out that there is a sense of “nonphysical” (namely nonphysicalistic) in which the CMoP of a phenomenal concept is indeed nonphysical. I noted, however, that physicalists are not committed to all language or conceptual apparatus being physicalistic.

Physicalists are committed to ontological physicalism, not conceptual reductionism. How does this apply to the MMoP-individuated-according-to-the-CMoP? It is true that if you individuate MMoPs according to CMoPs, then if there is no a priori available physical description, the MMoP is not “physical,” and, in this sense, White's argument is correct. But all “physical” comes to here is *physicalistic*, and it is no part of physicalism to make any claim to the effect that phenomenal MMoPs or CMoPs are physicalistic. Thus the assumption of the second argument (the topic of this section, the a priori availability argument), that the physicalist requires an a priori available description of the MMoP of the mental side of the mind-body identity, is false.

If MMoPs have to be thin, then perhaps the distinction between an MMoP's being ontologically physical and explicitly physical does not come to as much as would otherwise seem. Since a thin physical property has no hidden essence, it might be said to wear its physicality on its sleeve. However, if this is the only way to save the argument from a priori availability, the argument does not stand on its own but depends on the thin/thick argument, to which we now turn.

But first a brief reminder of what has been presupposed so far about the nature of MMoPs and CMoPs. In rebutting the regress argument, I assumed, along with the argument itself, that the *raison d'être* of MMoPs is to account for cognitive significance. The issue arose as to whether an MMoP defined according to its explanatory purpose must also fix reference or determine the referent. I noted that this cannot be assumed. The issue of the nature of CMoPs did not arise. In rebutting the second argument, I did not make any assumption about MMoPs or CMoPs that should be controversial, allowing a priori availability of the MMoP on the basis of the CMoP, reference-fixing authority, and determination of the referent.

Thin/Thick

The third argument that the MMoP of a phenomenal concept cannot be physical involves the distinction mentioned between “thin” and “thick” properties. As we have seen above, the first two parts of the Property Dualism Argument fall flat on end p.289

their own, but they can be resuscitated using the thin/thick distinction. However, if it could be shown that MMoPs must be thin, these other arguments would be superfluous, since the claim that MMoPs must be thin leads to dualism by a shorter route, as I will explain shortly.

First, I must consider what exactly the thick/thin distinction is. I have been taking it that whether a property is thick or thin is a matter of whether it has a hidden essence. On this view, the primary bearer of thickness is a property, and a thick concept would be a concept that purports to be a concept of a thick property. However, this definition will be wrong if fundamental physical properties are thin. For since being water = being H_2O , if being H_2O is thin and being water is thick, whether a property is thick or thin is relative to what concept one has of that property. (Of course, being H_2O is not a candidate for a fundamental physical property—I used that description as a surrogate because I don't know how to describe water in terms of electrons, quarks, etc.) On the picture of the thick/thin distinction in which whether a property is thin is concept-relative, one could define a thin concept as follows: the extension of the concept in a possible world does not depend on its extension in the actual world. (In terms of Chalmers's apparatus, the primary intension is the same as the secondary intension.) And thin properties would be defined in terms of their relation to thin concepts.³⁵

Are fundamental physical properties thin? Or, to put the matter from the other perspective, are fundamental physical concepts concepts of thin properties? We could approach the issue via the question of whether there could be a Twin Earth case for fundamental physical concepts. In my view, the answer is yes. I gave an example long ago (Block 1978) in terms of matter and antimatter, using for simplicity, the physics of the 1960s. The idea is that there is a counterfactual situation in which people who are relevantly like us—functionally like us—use the term “electron” to refer to antielectrons. That is, the counterfactual situation is one in which our doppelgängers inhabit a universe or a place in our universe in which antimatter plays the role played here by matter. And as a result, their Ramsey sentence for fundamental physics is the same as ours.³⁶ Which suggests that the functional role of a concept inside the head is not enough to determine its full nature, since the concept of an electron is not the same as the concept of an antielectron.

But perhaps science will delve further into the matter/antimatter distinction, coming up with further structure that explains the distinction and that would make a difference between the functional role of our concept and the doppelgängers' concept. The problem is that what we regard as fundamental physics is full of symmetries that can ground more complex examples, the idea being that there is more to physical reality than can be cashed out in a Ramsey sentence.

Of course, I don't think this mere suggestion settles the matter. Rather, I take the upshot to be that the issue of whether fundamental physical properties are thin cannot be settled here. Another argument in favor of that view is the point (Block

end p.290

2003) that it is compatible with much of modern physics that for each level, there is a still more fundamental level, the upshot being that there is something defective about the notion of a “fundamental” level in physics.

Ideally, I would consider the issues concerning the thick/thin distinction using both approaches, with thin properties defined in terms of thin concepts and vice versa. Here, however, I will simply make a choice based on ease of discussion: I'll take properties as basic. I don't think any issues will depend on this choice.

Whether a property is thick or thin, then, will be considered here to be a matter of whether it has a hidden essence. For example, water or the property of being water is thick, since whether something is water goes beyond superficial manifestations of it. Examples of thin properties are mathematical properties, at least some functional properties, and phenomenal properties if dualism is true. (The last point about dualism could be challenged—see Nagel 2002—but I will put the issue aside.) Artifact properties such as being a telephone might also be taken by some to be thin. As I mentioned, fundamental properties of physics might be alleged to be thin.

Note that it is not necessary for the property dualist to claim that *all* MMoPs are thin properties; it would be enough if this were true only for the MMoPs of phenomenal concepts. I do not have a blanket argument against all attempts to show that MMoPs for phenomenal concepts must be thin, but I do have arguments for a number of specific attempts.

Why believe that MMoPs must be thin? I will start with two arguments:

1. The a priority argument, which appeals to the idea that the MMoP is a priori available on the basis of the CMoP.
2. The aspect argument, according to which the cognitive significance role of MMoPs precludes thick properties serving as modes of presentation. For the property dualist may say that it is not *all* of a thick property but only an *aspect* of it, the thin aspect, that explains rational error.

These two arguments for MMoPs (at least for phenomenal concepts) being thin appeal to different features of MMoPs and their relations to CMoPs. Although I have registered doubt as to whether the same entities can serve both functions, I will put that doubt aside.

The A Priority Argument for Thin MMoPs

Let us assume that the MMoP of a concept is a priori available on the basis of the CMoP. For example, if one grasps the term “Hesperus,” and if its CMoP is the meaning or other mental features of “the morning star,” then the MMoP of rising in the morning is supposed to be a priori available by virtue of one's grasp of the term and its CMoP. This

constraint might be taken to rule out thick MMoPs, for it might be said that I do not know a priori whether I am on Earth or Twin Earth (McKinsey 1991). A thick MMoP might vary as between Earth and Twin Earth, which would be incompatible with a priori availability on the basis of the CMoP, which I and my twin on Twin Earth share.

I will give a fuller treatment of such arguments in the next section, but for now I will reply for the special case of phenomenal concepts, using the points made earlier about a phenomenal property doing double duty.

end p.291

I mentioned that a phenomenal property might be part of a CMoP, but also be brought in by the MMoP. For example, the CMoP might be taken to be the meaning or other mental features of: “the experience:___,” where the blank is filled by phenomenal property P. And the MMoP might be the property of being P. Such a relation between the CMoP and the MMoP allows for the MMoP to be a priori available on the basis of the CMoP, even if the property P is a thick property with a hidden essence. That is, the property of *being* P is a priori available on the basis of a grasp of a CMoP that has property P as a constituent whether or not P is thick.

Although the a priori relation in itself does not appear to pose an obstacle to the thickness of the MMoP, it might be thought to pose a problem combined with another argument, to which we now turn.

The Aspect Argument

As mentioned, the idea of the aspect argument is that it is not *all* of a thick property that explains cognitive significance in general and rational error in particular, but only an aspect of it, the aspect that is available a priori on the basis of the CMoP. But on the face of it, that aspect can itself be thick. Recall the example of Albert, which I pick out on the basis of its being the local wet thing. Albert's property of being the local wet thing fixes reference, uniquely determines the referent, is a priori available, and is thick.

The property dualist may say that the property that would serve in explanations of error is not that it is wet but that it *looks* wet. However, consider a nonperceptual case: I infer, using inductive principles, that something in the corner is wet, and I pick it out via its property of being wet. In this case, the substitution of *looks wet* for *wet* is unmotivated.

The MMoP just does not seem perceptual. Nor does it seem artifactual nor, more generally, functional. On the face of it, the MMoP is a thick property, the property of being wet—that is, at least partially covered or soaked with water (which is thick because being covered or soaked with water is being covered or soaked with H₂O).

But perhaps this rebuttal misses the significance of aspects to the first-person point of view. Perhaps the property dualist will say something like this:

If phenomenal property Q is a physical property, then it can be picked out by a physical—say, neurological—concept that identifies it in neurological terms. But those neurological identifications are irrelevant to first-person phenomenal identifications, showing that the first-person phenomenal identification depends on *one aspect* of the

phenomenal property (its “feel”) rather than *another aspect* (its neurologically identifying parameters). You have suggested that “cortico-thalamic oscillation” picks out its referent via the effect of cortico-thalamic oscillation on instruments that monitor oxygen uptake from blood vessels in the brain. But this effect is not part of the first-person route by which we pick out Q, so it follows that not every aspect of the physical property is relevant to the first-person route. Therefore the identity “Q = cortico-thalamic oscillation” is supposed to be one in which the terms pick out a single referent via different properties of it, different MMoPs. And so the Property Dualism Argument has not been avoided. I agree that the two terms of the identity “Q = cortico-thalamic oscillation” pick out the referents via different aspects of that referent, different MMoPs. And I also
end p.292

agree that the aspect used by the mental term of the identity is available to the first person whereas the aspect used by the physical term is not. But it does not follow that the aspect used by the mental term is thin. It is true that no neurological property is explicitly part of the first-person route, but that does not show that it is not part of the first-person route, albeit ontologically rather than explicitly. The MMoP of Q is stipulated to be phenomenal, and may be taken to be the property of *being* Q. But being identical to Q, on the physicalist view, is *both* a thick property and available to the first person. Being identical to Q is a physical property (being identical to cortico-thalamic oscillation) but is nonetheless distinct from the MMoP I have been supposing for “cortico-thalamic oscillation,” which has to do with the oxygen uptake that functional magnetic resonance scanners use to identify it. On the physicalist view, the feel and the neurological state are not different aspects of one thing: they are literally identical. If they are aspects, they are identical aspects. But the MMoP of the right-hand term of the identity is still different from the MMoP of the left-hand term.

As mentioned earlier, some will say that oxygen uptake cannot provide the MMoP for the term “cortico-thalamic oscillation,” which should be taken to be cortico-thalamic oscillation itself, or perhaps being identical to cortico-thalamic oscillation. In this supposition, there is a germ of a different argument for dualism, the E → 2M Argument discussed earlier.

I say that the aspect of a property that accounts for cognitive significance can itself be thick, appealing to examples. But the property dualist may suppose that if we attend to the mental contents that are doing the explaining, we can see that they are *narrow* contents, contents that are shared by Putnamian twins, people who are the same in physical properties inside the skin that are not individuated by relations to things outside the skin. If the relevant explanatory contents are narrow contents, then the corresponding explanatory properties—MMoPs—will be thin.

Here is the argument, the Narrow → Thin Argument, in more detail, offered in the voice of the property dualist:

N → T Argument: Suppose my CMoP is “the wet thing in the corner” (in a nonperceptual case), and my twin on Putnam's Twin Earth would put his CMoP in the same words. Still, the difference between what he means by “wet” and what I mean by “wet” *cannot matter to the rationalizing explanatory force* of the CMoPs. And since CMoPs are to be individuated entirely by rationalizing explanatory force, my twin and I

have the same CMoPs: the CMoPs are narrow. But since the MMoP is a priori available to anyone who grasps the CMoP, the twins must have the same MMoP as well as the same CMoP, so the MMoP must be thin. Narrow CMoP, therefore thin MMoP.

The $N \rightarrow T$ Argument presupposes the familiar but controversial idea that only narrow content can serve in intentional explanations. However, on the face of it, my “water” concept can be used in an explanation of my drinking water (“I wanted water, I saw water, so I drank water”) but would not explain my drinking twin-water.³⁷ The idea that only narrow contents can serve in a rationalizing explanation is certainly controversial. I will not enter into this familiar dispute here except to
end p.293

say that I think the papers by Burge referred to in the last footnote make a convincing case for wide explanations.

The inference from narrow content/narrow CMoP to thin MMoP has some initial plausibility, but it in fact begs the question. I agree with the premise of the $N \rightarrow T$ Argument that phenomenal CMoPs are narrow. (I won't go into the possibility that there is a descriptive part of the CMoP that is wide.) However, it does not follow that the MMoP is thin. The physicalist says that since phenomenality supervenes on the physical, Putnamian doppelgängers will share CMoPs: CMoPs are narrow. For example, a phenomenal CMoP containing phenomenal property P for one twin will also contain phenomenal property P for the other twin. The MMoP, being P, will also be the same for both twins, but that MMoP can nonetheless be thick. In short, the phenomenal part of a CMoP and the corresponding phenomenal MMoP will in general be narrow in virtue of being necessarily shared by doppelgängers, but will nonetheless be thick on the physicalist view. That is, what the doppelgängers necessarily share will be a property with a scientific essence.

The point can be approached by looking at the anomalous nature of phenomenal kinds. Phenomenal concepts of the sort that I have described here are natural kind concepts in that they purport to pick out objective kinds, and if the physicalist is right, those kinds have scientific natures whose scientific descriptions cannot be grasped a priori simply on the basis of having the concept. But they differ from most natural kind concepts in that the Twin Earth mode of thought experiment does not apply. The Twin Earth mode of thought experiment involves a pair of people who are the same in physical properties inside the skin (that are not individuated by relations to things outside the skin) but with a crucial physical difference. In Putnam's classic version, twins who are relevantly the same in physical properties inside the skin pick out substances using the term “water” that have physically different natures, so (it is claimed) the meanings of their “water” terms and “water” thought contents differ. They are (relevantly) physically the same, but different in “water” meaning and “water” content.

But how is the Twin Earth thought experiment supposed to be applied to phenomenality? If physicalism is true, the twins cannot be the same in physical properties inside the skin (that are not individuated by relations to things outside the skin) and also differ in the physical natures of their phenomenal states! (That's why I say the $N \rightarrow T$ Argument begs the question against physicalism.) So there is no straightforward way to apply the

Putnamian Twin Earth thought experiment to phenomenal concepts. (The issue concerning Burgean thought experiments is more complex because it hinges on the ways our terms express phenomenal concepts. I can't go into the matter here.)

But perhaps only a superficial analysis of Twin Earth thought experiments requires that the twins be the same in physical properties inside the skin (that are not individuated by relations to things outside the skin). One way to think of Twin Earth cases is that what is important is that they be *mentally* alike in ways that don't involve relations to things outside the skin. (Thus, for some purposes, functional likeness may seem more relevant than microphysical likeness. This line of thought was what I used in my earlier discussion of whether fundamental physical properties are thin.) But phenomenality is certainly part of mentality, so if twins are

end p.294

to be the same in phenomenal CMoPs, there had better not be any physical difference between them that makes a phenomenal difference. However, from the physicalist point of view, the shared phenomenality of the twins' CMoPs has to be explained by a shared physical basis of it. So the shared narrow CMoP is compatible with a shared thick MMoP.

The upshot is that phenomenal concepts are an *anomaly*—at least from the physicalist point of view. They are natural kind concepts in that they allow for objective scientific natures that are “hidden” (the scientific descriptions are not a priori available on the basis of merely having the concept). But they are different from other natural kind concepts in that no reasonable facsimile of a Putnamian Twin Earth scenario is possible.

So even if the inference from narrow CMoP to thin MMoP applies in a variety of other cases, it should not be surprising that it fails to apply in this anomalous case. The CMoPs for phenomenal concepts can be narrow even though the corresponding MMoPs are thick. Indeed, the CMoPs themselves can be both narrow *and* thick. Narrow because nonrelational in the appropriate way, thick because they involve a phenomenal element that has a hidden scientific nature.

I have rebutted the aspect and a priority arguments and a subsidiary argument, the $N \rightarrow T$ Argument, which all push for the conclusion that MMoPs of phenomenal concepts must be thin. But one can also look at the thesis itself independently of the arguments for it. Here are two considerations about the thesis itself.

Issues about the Claim of Thin MMoPs for Phenomenal Concepts

First, the assumption of thin MMoPs is perhaps sufficient for the conclusion of the Property Dualism Argument *all by itself*. For what are the candidates for a thin MMoP for a phenomenal concept? Artifact properties like being a telephone (even assuming that they are thin) and purely mathematical properties are nonstarters. Some kind terms that are not natural kind terms (e.g., “dirt”) may yield thin properties. But phenomenal MMoPs are not artifactual or mathematical, and they are or purport to be natural kinds. It is not clear whether there are any natural kind terms that express thin properties. Even if

there are fundamental physical properties that are thin, the property dualist can hardly suggest fundamental physical properties as candidates for MMoPs for phenomenal concepts, since that has no independent plausibility and in any case would be incompatible with the conclusion of the property dualist's argument. So it would seem that the only remotely plausible candidates for thin MMoPs by which phenomenal concepts refer are (1) purely functional properties, in which case deflationism would be true, and (2) phenomenal properties that are nonphysical, in which case dualism is true. The conclusion would be the same as the conclusion of the Property Dualism Argument itself: that phenomenal realist physicalism is untenable.

The upshot is that much of the argumentation surrounding the property dualism argument can be dispensed with if the arguments of this chapter are correct. The most obvious arguments that MMoPs of phenomenal concepts cannot be physical (the regress argument and the a priori availability argument) do not stand alone but rather depend on the thin/thick argument. I have not shown that there is no good
end p.295

argument for the claim that MMoPs of phenomenal concepts are thin, but I have rebutted some obvious candidates, and it is hard to see how the regress and a priori availability arguments could be used to justify the thinness claim, given that they presuppose it. So if my arguments are right, the burden of proof is on the property dualist to come up with a new argument for the claim that MMoPs of phenomenal concepts are thin.

Here is the second point. So far, I have argued that the assumption of thin MMoPs leads directly to dualism or deflationism, putting a heavy burden of proof on the property dualist to justify that assumption. But actually I doubt that deflationism really is an option. Let me explain. The functionalist characterizes functional properties in terms of the "Ramsey sentence" for a theory. Supposing that "yellow teeth" is an "observation term," the Ramsey sentence for the theory that smoking causes both cancer and yellow teeth is $\exists F_1 \exists F_2 [F_1 \text{ causes both } F_2 \text{ and yellow teeth}]$; the Ramsey sentence says that there are two properties one of which causes the other and also yellow teeth. Focusing on psychological theories, where the "observation terms" (or "old" terms in Lewis's parlance) are terms for inputs and outputs, the Ramsey sentence could be put as follow: $\exists F_1 \dots \exists F_n [T(F_1 \dots F_n, \mathbf{i}_1 \dots \mathbf{i}_m, \mathbf{o}_1 \dots \mathbf{o}_p)]$. The "i" terms are input terms and the "o" terms are output terms. Functional properties of the sort that can be defined in terms of the Ramsey sentence are properties that consist in having certain other properties that have certain causal relations to inputs, outputs, and other properties.³⁸ The inputs and outputs can be characterized in many ways. For example, an output might be characterized neurally, or in terms of movements of a hand or leg, or distally, in terms of, for example, water in the distance, or distally and mentalistically in terms of drinking water. *But all of these characterizations are plausibly thick, not thin.* Perhaps you will think that some of them are *themselves* to be cashed functionally, but then the issue I am raising would arise for the input and output specification of *those* functional properties. Since the problem I am raising depends on the thickness of the input and output properties, I put those terms for those properties ("**i**₁" ... "**i**_m," "**o**₁" ... "**o**_p") in bold in the Ramsey sentence earlier. The only functional properties I know of that are plausibly thin are *purely formal* functional properties that abstract from the specific

nature of inputs and outputs, the kind of functional property that could be shared by a person and an economy. (See Block 1978.) For example, in the case of the theory that smoking causes cancer and yellow teeth, a purely formal Ramsey property would be: being an x such that $\exists F_1 \exists F_2 \exists F_3 [F_1 \text{ causes } F_2 \text{ and } F_3]$ and x has F_1 . This is the property of having a property, which causes two other properties. Such a property could be shared by a person and an economy. Since not even a deflationist should agree that the metaphysical modes of presentation of our phenomenal states are *purely* formal, the only remaining option is dualism. So the assumption of thin properties plausibly leads right to dualism.

end p.296

To sum up the points about the thin/thick argument: The “aspect” rationale for MMoPs being thin seems doubtful because the aspect can itself be thick. And the rationale for thin MMoPs in terms of the supposed a priori relation between CMoP and MMoP is problematic because the key phenomenal feature of the MMoP can also be present in the CMoP when the relevant concept is phenomenal. At least this is so on one plausible notion of phenomenal concepts, which the property dualist would have to challenge. Narrow CMoPs can be used to argue for thin MMoPs, but this reasoning begs the question against the physicalist. I explained at the outset that the emphasis on MMoPs at the expense of CMoPs was tactical: the metaphenomenal move, which says that modes of presentation bring in unreducible phenomenality, can be discussed equally with respect to either mode of presentation. This is the place in the argument where the artificiality is most apparent; CMoPs must be discussed explicitly.

Moving to the thesis itself, independently of arguments for it, the assumption of thin MMoPs amounts to much the same thing as the Property Dualism Argument itself. Further, the only remotely plausible candidates for thin MMoPs are purely formal properties that we do not have ordinary concepts of and phenomenal properties dualistically conceived. The purely formal properties, though more plausible than some other candidates, are not very plausible, even from a deflationist point of view. Deflationist functionalism is based on analyses of mentality in terms of sensory input and behavioral output. Purely formal properties do not adequately capture such analyses and cannot do so without thick input and output terms. The upshot is that the assumption of thin MMoPs for phenomenal concepts adds up to dualism itself. To assume thin MMoPs begs the question against the physicalist.

Of course, I have not shown that there cannot be an argument for thin phenomenal MMoPs, but I hope I have shown that a number of candidates do not succeed.

The Relation between the Property Dualism Argument and Some Other Arguments for Dualism

Loar (1990/97) locates the flaw in Jackson's “Mary” argument and Kripke's modal argument in a certain principle, the “semantic premise.”³⁹ The semantic premise (on one understanding of it) says that if a statement of property identity is a posteriori, then at

least one of the MMoPs must be contingently associated with the referent. The idea behind the principle is that if the two concepts pick out a property noncontingently, it must be possible for a thinker who grasps the concepts to see, a priori, that they pick out the same property. Again the issue arises as to what notion of MMoP is at stake. Consider, for example, the reference-fixing notion of MMoP. In this sense, the “semantic premise” is plainly false. Note that the person formed by
end p.297

a certain sperm = the person formed by a certain egg. This identity is a posteriori, yet both terms pick out their referents via essential and therefore necessary properties of it, assuming that Kripke is right about the necessity of origins. Call the sperm and egg that formed George W. Bush “Gamete-Herbert” and “Gamete-Barbara,” respectively. The person formed from Gamete-Herbert = the person formed from Gamete-Barbara. “The person formed from Gamete-Herbert” does not pick out George W. contingently, nor does “The person formed from Gamete-Barbara.” My example is put in terms of individuals but it is easy to see how to frame a version of it in terms of properties. Even if Kripke is wrong about the necessity of origins, the logic of the example remains. One thing can have two necessary but insufficient property, both of which can be used to pick it out, neither of which a priori entails the other. Thus the terms in a true a posteriori identity can pick out that thing, each term referring by a different necessary property as the MMoP.

Of course there is some contingency in the vicinity. Gamete-Herbert might have joined with an egg other than Gamete-Barbara or Gamete-Barbara might have joined with a sperm other than Gamete-Herbert. And this might suggest a modification of the principle (one that White [chap. 11, this volume] suggests in response to an earlier version of this chapter), namely, that if a statement of property identity is a posteriori, then it is not the case that both terms refer via MMoPs that are necessary and sufficient conditions for the property that is the referent. Or, more minimally, if a property identity is a posteriori, then it is not the case that one term refers via a sufficient property of it and the other refers via a necessary property of it. But a modification of my example (contributed by John Hawthorne) suggests that neither of these will quite do. Let the identity be the *actual* person formed from Gamete-Herbert = the *actual* person formed from Gamete-Barbara. Arguably, each designator refers via a property that is both necessary and sufficient for the referent. So the revised version of the semantic premise is also false. The reference of the terms “Gamete-Herbert” and “Gamete-Barbara” need not be fixed via properties that involve George Bush. The gametes can be identified independently, for example, before George Bush was conceived. But perhaps the names will pick them out via some contingent reference-fixing property, such as a perceptual demonstrative (“that egg”) or by description. And that motivates White (chap. 11, this volume) to suggest a beefed-up form of the semantic premise that says that there must be contingency either in the relation between MMoPs and referent or in the relation between MMoPs and the MMoPs of those MMoPs, or ... , and so on. I reject the beefed-up semantic premise for the reason given earlier: I don't think these further MMoPs need exist. That is, in the identity “a = b,” there will be MMoPs associated with both sides. But there will be no MMoPs of those MMoPs unless the subject happens to refer to the first-level MMoPs in another voluntary cognitive act.

Conclusion

Both the Knowledge Argument and Max Black's Property Dualism Argument for dualism hinge on the idea that there is something special about phenomenality in end p.298

our phenomenal concepts that eludes physicalism. Both arguments are ways of making concrete what I called the metaphenomenal move: the idea that in a phenomenal mind-body identity claim, the CMoP is partly constituted by something with unreducible phenomenality or the MMoP is an unreducible phenomenal property.

My response has been to argue that phenomenality in modes of presentation is no different from phenomenality elsewhere. I tried to dissolve apparent impediments to the phenomenal element in the CMoP and the MMoP being physical. My way out involves a notion of a phenomenal concept that has some affinities with the “directness” story in which there is no metaphysical mode of presentation at all, since my phenomenal MMoPs are not very different from the referent itself. I considered a family of arguments based on the idea that MMoPs must be thin, arguing that appeal to narrow content does nothing to establish thinness. My own view is that phenomenal concepts are both narrow and thick, which is why the phenomenality in the CMoP can be physical. I also considered a version of the Property Dualism Argument which assumes that an empirical identity must have different MMoPs, so that if the MMoPs of the two terms of an identity are the same, then the identity is a priori. I argued that whereas sameness of CMoP makes for a priority of the identity, sameness of the MMoP does not.

Much of the argumentation involved the principle that a difference in CMoP requires a difference in MMoP ($D(\text{CMoP}) \rightarrow D(\text{MMoP})$). I argued that nothing forces us to adopt notions of CMoP and MMoP on which this principle is true. However, at a key point in the dialectic, I considered a notion of MMoP individuated with respect to CMoP, which I argued did not rescue the Property Dualism Argument.

Although I expressed skepticism about whether any one thing can explain rational error, fix reference, and be relevantly a priori available, I have not claimed that these *raison d'être* fail to coincide except at one point at which I noted that an explanatory MMoP need not fix the referent. The other rebuttals were keyed to one or another specific version of MMoPs and CMoPs and their relation. My strategy was to avoid multiplying arguments based on different notions of CMoP/MMoP by choosing what seemed to me the strongest argument of each type. In the end, everything hinges on the claim that MMoPs of phenomenal concepts are thin, and I attempted to remove the most straightforward motivations for that view.

I have pursued a divide-and-conquer strategy, distinguishing among different senses of “mode of presentation” and further dividing those by the *raison d'être* of modes of presentation in those senses. My claim is that once we do that, the Property Dualism Argument dribbles away. I have not claimed to conclusively refute these arguments, but I believe that the ball is in the property dualist's court.

Appendix on a Variant of the $E \rightarrow 2M$ Argument Using Primary Intensions Instead of MMoPs

I mentioned that there is a version of the $E \rightarrow 2M$ Argument that uses the notion of a primary intension instead of the notion of an MMoP. The primary intension of “water” is the function from worlds considered as actual (i.e. as actual world
end p.299

candidates in Davies and Humberstone's sense) to what water turns out to be in that world. (Or so I will understand the term. Chalmers uses various ways of specifying what a primary intension is, but since this is a very brief discussion, I will just pick that one.) Thus the primary intension of “water” picks out water in the actual world and XYZ (“twin-water”) on Putnam's Twin Earth. Since Putnam's Twin Earth could have both XYZ and H_2O in it, the primary intension is a function from “centered” worlds—worlds with a privileged point—to referents. What makes the primary intension of “water” pick out XYZ in Putnam's Twin Earth is that the center of that world has the relevant relation to XYZ rather than to H_2O . (For example, the center might be surrounded by XYZ, whereas there might be only a few molecules of H_2O that are light years away and that have not causally impinged on the center. If there are people on Twin Earth, we can suppose causal commerce of the relevant sort between XYZ [but not H_2O] and uses of the term “water” that have some appropriate relation to the center.)

I read Chalmers (1996) as stipulating that the primary intension captures the a priori component of content, and on this reading the primary intension would be more like a CMoP than an MMoP. Given this stipulation, my complaint that MMoPs are not what is relevant to a priority would fall away, the pressure instead being on the issue of whether primary intensions in the sense in which they are stipulated to capture the a priori aspect of content are indeed the same as secondary intensions. That is, the analog of the $E \rightarrow 2M$ principle, the “ $E \rightarrow 2PI$ principle,” would be a stipulation, but the other premise of the argument—that the phenomenal and functional primary intensions are identical to the secondary intensions—would then take the heat. The secondary intension of “water” is the function from worlds to what “water” denotes in those worlds, namely water, if there is any, i.e. H_2O . The worlds are considered “as counterfactual” (as is familiar from Kripke): we take the reference of “water” in the actual world as fixed, and, given that fixed reference, the function picks out what is identical to the actual referent in each counterfactual world, namely H_2O if there is any (assuming the usual philosophical myth that “water” refers to H_2O in the actual world). So the doubtful premise—according to me—would be whether primary intensions that are stipulated to capture the a priori component of content are the same as secondary intensions for either the phenomenal or the psychofunctional term. (I will be explaining why I am doubtful that primary intensions can be stipulated to capture an a priori component of content.) Of course, there is no plausibility of “primary intension = secondary intension” for “water.” Twin Earth is a counterexample because the primary intension picks out XYZ, whereas the secondary intension picks out H_2O . To the extent that the right-hand side of mind-body identity claims are natural kind terms like “water,” the version of the Property Dualism Argument presented in this appendix has no plausibility whatsoever. You can see why by noting that primary intensions in this incarnation correspond to my CMoPs—

which can also be stipulated to capture the a priori aspect of content. As noted earlier, there is no plausibility at all that the CMoP of, say, a functional term, is identical to the referent. Consider a very simple functional term, “solubility.” The CMoP of “solubility” is something like a meaning, but the referent is a property of sugar and salt. Why should we suppose
end p.300

that the solubility of sugar and salt is a kind of *meaning*? This seems to be a category error.

In many of his writings, Chalmers has one notion—primary intension—corresponding to the two notions of my apparatus—CMoP and MMoP. But in Chalmers 2006, he considers dividing the primary intension into two notions. As I understand it, the *epistemic* intension of “water,” which is stipulated to capture the a priori aspect of content, is a function from situations (not worlds) to what turns out to be water in those situations. The primary intension, which on this version is not stipulated to capture the a priori aspect of content, is a function from worlds to what turns out to be water in those worlds. So on this scheme, epistemic intensions roughly correspond to my CMoPs, whereas primary intensions roughly correspond to MMoPs. On this new notion of a primary intension, it becomes a substantive question whether primary intensions capture an a priori component of content. If it turns out that they do for phenomenal terms and psychofunctional terms, then the Property Dualism Argument would avoid the first of the two objections I mentioned above to the $E \rightarrow 2M$ analog for primary intensions. So it is worth taking a closer look at the prospects for the substantive (as opposed to stipulated) claim that primary intensions capture an a priori component of content.

I will criticize the primary intension as stipulated to capture an a priori component of content (see Block 1991, and Block and Stalnaker 1999; see also Chalmers 2006 for a response). Take the value of the primary intension of “water” to be what turns out to be water in a world considered as actual. How do we know what turns out to be water in a world considered as actual? By consulting our intuitions about what one should say about various worlds considered as actual. We ask ourselves what we should say if, for example, we became convinced we were living in Putnam's Twin Earth. These intuitions are the epistemic basis of the primary intension, that is how we know what it is. And they are or at least index its metaphysical basis. That is, these intuitions constitute the metaphysical basis or they index an underlying property that is responsible both for the intuitions and the primary intension.

Now ask yourself about another of Putnam's (1970) thought experiments, which we could put like this. Suppose we discover that cats are actually robots controlled from Mars that were put on Earth 100 million years ago to spy on the intelligent beings they predicted would evolve. There never were any naturally evolving catlike creatures, since the robot cats killed off anything that had a chance of becoming one. When intelligent primates finally evolved, the robot cats made themselves appealing to people and came to develop the close relation to people portrayed in *Garfield*. We are wrong about many of the properties we take cats to have. The robot cats pretend to be aloof but are actually very interested in us and love us. They would like nothing better than to act more like dogs,

but their orders are to act aloof. They do not actually purr but use mind control to make us think they are doing so.

I think the story is intelligible, and I hope you think so, too. But notice that there might be other (equally intelligible) stories in which cats also fail to have other properties that we ordinarily think they have. The world I mentioned is a world considered as actual in which cats are not cute, aloof, purring animals. But there are other worlds considered as actual in which they lack other properties that we ordinarily think they have. Perhaps *all* the properties we ascribe to cats—or at
end p.301

least the ones that distinguish them from, say dogs—are in this sense dispensable. Some may want to retreat to *seeming* to have such properties, but in this direction lies the phenomenalism of C. I. Lewis. If the primary intension of “cat” is determined or indexed by such intuitions and captures the a priori component of content, it would appear that there is very little to the a priori component of content. Maybe one can't imagine a world considered as actual in which cats are not moving, middle-size physical entities—but that will not distinguish a putative a priori component of “cat” from that of “dog.”

The Chalmers-Jackson response is to note that our intuitions about worlds considered as actual do in fact distinguish between “cat” and “dog,” so the primary intensions are not so thin as to be the same for these two words. This response, however, sets up the real worry, which is that given that these intuitions are or at least index the foundation of the semantics of these terms, how we are supposed to know whether, in having these intuitions, that is, in considering a world as actual, we end up *covertly changing the meaning of “cat.”* That is, how do we know whether in coming up with the best way of thinking about a world as actual, one of the variables we implicitly adjust is the meanings of the words we use to describe the world?

The problem would be avoided if one had some *other* notion of the a priori component of content that could be used in defining primary intensions, for example an account along the lines of the suggestion from Kripke that some words can be defined metalinguistically or, alternatively, Katz's more orthodox definitions. The primary intension of “cat” would be the function from worlds considered as actual to what is picked out in that world by the proposed definition. But then we would not need the primary intension as an account of the a priori component of content because we would already have such an account: the definition.

Note that the problem is not one of indeterminacy in our intuitions or of cases not decided by our intuitions. Of course, there are cases our intuitions do not decide. The problem is with cases that our intuitions *do* decide, such as the robot cat case. Our intuitions are a function of the simplest overall account, and as Quineans have long said, there is no guarantee that anything putatively a priori will be preserved in the simplest account. If one believes in determinate a priori intensions, the thing to say is that our intuitions present us with situations in which we find it natural to change those a priori intensions. That is, in considering the Putnam robot cat world, we tacitly change our meaning of “cat” (Katz 1972, 1975).

So there is a dilemma for the advocate of primary intensions as stipulated to capture an a priori component of content. If our advocate goes with Katzian or metalinguistic

definitions, then there is no need for the notion of a primary intension. However, if our advocate rejects those definitions, then it is not clear why we should believe that our linguistic intuitions index or determine any interesting a priori aspect of content or the primary intensions that are stipulated to capture it. Of course, primary intensions are just functions, and so the primary intension of “cat” can be said to exist trivially. Yes, but that function may include inputs in which the word “cat” is used in a different sense from the normal one and so could not be said to capture anything semantic. (See the coumarone example in Block and Stalnaker 1999.) The question is: Why should we believe in a primary intension that does capture an a priori aspect of content? Given the unreliability of the intuitions about
end p.302

cases as a pipeline to an a priori notion of content, primary intensions which are stipulated to capture an a priori notion of content become highly doubtful theoretical entities.

The upshot for the analog of the $E \rightarrow 2M$ form of the Property Dualism Argument is this. If an intension, primary or epistemic, is simply stipulated to capture an a priori aspect of content, then it is in doubt for the reasons just given. If we put this doubt aside, accepting the analog of the $E \rightarrow 2M$ principle, the identity of those intensions with secondary intensions is in doubt—that is, the other premise of the argument is in doubt. What if the intension is not stipulated to have this a priori significance, but it is claimed to have it nonetheless? The Putnamian considerations I raised cast doubt on that claim, but putting that doubt aside, my view is that to the extent we can show an intension to capture an a priori aspect of content, it will be doubtful that that intension can be identified with a secondary intension, so the two premises of the $E \rightarrow 2M$ argument cannot be satisfied together.

Acknowledgments

I thank the following persons for commenting on a remote ancestor of this chapter: Paul Horwich, Brian Loar, David Pitt, Stephen Schiffer, Susanna Siegel, Stephen White, and Dean Zimmerman. Thanks also to Tyler Burge, David Chalmers, and Stephen White for comments on a more recent version. I am grateful to students in my graduate seminar, participants in the NEH Santa Cruz Summer Institute of 2002, participants at an Australian National University Workshop (“Themes from Ned Block”) in the summer of 2003, and an audience at the University of Houston for reactions to parts of the remote ancestor.

This chapter is reprinted with permission from D. Zimmerman, ed., *Oxford Studies in Metaphysics*, Volume 2. New York: Oxford University Press, 2006: 3–78.

References

Armstrong, D. (1968) *A Materialist Theory of the Mind*. London: Routledge.

Austin, D. F. (1990). *What's the Meaning of "This"?* Ithaca, N.Y.: Cornell University Press.

Balog, K. (2006) Acquaintance and the Mind-Body Problem. On the web at <http://pantheon.yale.edu/%7Ekb237/Web%20publications/Acquaintance.pdf>

Block, N. (1978). Troubles with Functionalism. In *Perception and Cognition. Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science*, Vol. 9, ed. W. Savage: 261–325. Minneapolis: University of Minnesota Press. Reprinted in *Readings in Philosophy of Psychology*, Vol. 1, ed. N. Block: 268–305. Cambridge: Harvard University Press, 1980. Shortened version in *Mind and Cognition*, ed. W. Lycan: 444–69. Oxford: Blackwell, 1990. Reprint of shortened version in *The Nature of Mind*, ed. D. Rosenthal: 211–29. Oxford: Oxford University Press, 1991. Revised shortened version in *Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 94–98. Oxford: Oxford University Press, 2002.

Block, N. (1980). What Is Functionalism? In *Readings in Philosophy of Psychology*, Vol. 1, ed. N. Block: 171–84. Cambridge: Harvard University Press.

Block, N. (1991). What Narrow Content Is Not. In *Meaning in Mind: Fodor and His Critics*, ed. B. Loewer: 33–64. Cambridge: Blackwell.

end p.303

Block, N. (1992). Begging the Question against Phenomenal Consciousness. *The Behavioral and Brain Sciences* 15: 205–06. Reprinted in *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzeldere: 175–79. Cambridge: MIT Press, 1997.

Block, N. (1993). Review of Daniel Dennett. *Consciousness Explained*. *Journal of Philosophy* 90: 181–93.

Block, N. (1994). Functionalism. In *A Companion to Philosophy of Mind*, ed. S. Guttenplan: 323–33. Oxford: Blackwell.

Block, N. (2002). The Harder Problem of Consciousness. *Journal of Philosophy* 99: 1–35. A longer version is in *Disputatio* 15 (November 2003).

Block, N. (2003). Do Causal Powers Drain Away?. *Philosophy and Phenomenological Research* 67: 110–27.

Block, N., and Dworkin, G. (1974). IQ, Heritability, and Inequality, Part 1. *Philosophy and Public Affairs* 3: 331–409.

Block, N., and Stalnaker, R. (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *Philosophical Review* 108: 1–46. [Link ▶](#)

Block, N., Flanagan, O., and Güzeldere, G. (eds). (1997). *The Nature of Consciousness*. Cambridge: MIT Press.

Boyd, R. (1980). Materialism without Reductionism: What Physicalism Does Not Entail. In *Readings in Philosophy of Psychology*, Vol. 1, ed. N. Block: 67–106. Cambridge: Harvard University Press.

Burge, T. (1977). Belie. *De Re*. *Journal of Philosophy* 74: 338–62. [Link ▶](#)

Burge, T. (1982). Two Thought Experiments Reviewed. *Notre Dame Journal of Formal Logic* 23: 284–93.

Burge, T. (1986). Individualism and Psychology. *Philosophical Review* 95: 3–45.

[Link ▶](#)

Burge, T. (1989). Individuation and Causation in Psychology. *Pacific Philosophical Quarterly* 70: 303–22.

Burge, T. (1995). Intentional Properties and Causation. In *Philosophy of Psychology: Debates about Psychological Explanation*, ed. C. Macdonald and G. Macdonald: 226–35. Oxford: Blackwell.

Byrne, A., and Pryor, J. (2006). Bad Intentions. In *The Two-Dimensionalist Framework: Foundations and Applications*, ed. M. Garcia-Carpintero and J. Macia.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.

Chalmers, D. J. (2003). The Content and Epistemology of Phenomenal Belief. In *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic: 220–72. Oxford: Oxford University Press.

Chalmers, D. J. (2004). Phenomenal Concepts and the Knowledge Argument. In *There's Something about Mary*, ed. P. Ludlow, D. Stoljar, and Y. Nagasawa: 269–98. Cambridge: MIT Press.

Chalmers, D. J. (2006). The Foundations of 2D Semantics. In *Two-Dimensional Semantics: Foundations and Applications*, ed. M. Garcia-Carpintero and J. Macia. New York: Oxford University Press. Abridged version, Epistemic Two-Dimensional Semantics, was published in *Philosophical Studies* 118, 1–2 (2004): 153–226.

Chomsky, N. (2000). *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.

Davies, M., and Humberstone, L. (1980). Two Notions of Necessity. *Philosophical Studies* 38: 1–30. [Link ▶](#)

Devitt, M. (1981). *Designation*. New York: Columbia University Press.

end p.304

Feigl, H. (1967). *The “Mental” and the “Physical”; the Essay and a Postscript*. Minneapolis: University of Minnesota Press. Originally published in *Minnesota Studies in Philosophy of Science* 3 (1958): 370–497.

Field, H. (1994). Deflationist Views of Meaning and Content. *Mind* 103: 249–85. [Link ▶ OUP Resource](#)

Fodor, J. (1982). Cognitive Science and the Twin-Earth Problem. *Notre Dame Journal of Formal Logic*, 23: 98–119. [Link ▶](#)

Hempel, C. (1969). Reduction: Ontological and Linguistic Facets. In *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel*, ed. S. Morgenbesser, P. Suppes, and M. White: 179–99. New York: St. Martin's Press.

Hill, C. (1991). *Sensations: A Defense of Type Materialism*. Cambridge: Cambridge University Press.

Hill, C. (1997). Imaginability, Conceivability, Possibility, and the Mind-Body Problem. *Philosophical Studies* 87: 61–85. [Link ▶](#)

Horgan, T. (1984). Jackson on Physical Information and Qualia. *Philosophical Quarterly* 34: 147–83. [Link ▶](#)

Horwich, P. (1990/98). *Truth*. Oxford: Blackwell.

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36.

 [Link](#)

Jackson, F. (1986). What Mary Didn't Know. *Journal of Philosophy* 83: 291–95.

 [Link](#)

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford: Oxford University Press.

Jackson, F. (2004). Review of John Perry. *Knowledge, Possibility, and Consciousness*. *Mind* 113: 207–10.

Katz, J. (1972). *Semantic Theory*. New York: Harper and Row.

Katz, J. (1975). Logic and Language: An Examination of Recent Criticisms of Intentionalism. In *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science* 7, ed. K. Gunderson: 36–130. Minneapolis: University of Minnesota Press.

Kripke, S. (1972/1980). Naming and Necessity. In *The Semantics of Natural Language*, ed. G. Harman and D. Davidson, 253–355. Dordrecht: Reidel. Reprinted with a Preface and corrections as *Naming and Necessity*. Cambridge: Harvard University Press, 1980.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press.

 [Link](#) 

Lewis, D. (1980). Mad Pain and Martian Pain. In *Readings in the Philosophy of Psychology*, Vol. 1, ed. N. Block: 216–22. Cambridge: Harvard University Press.

Loar, B. (1988). Social Content and Psychological Content. In *Contents of Thoughts*, ed. R. Grimm and P. Merrill: 99–110. Tucson: University of Arizona Press.

Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives* 4: *Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview. Revised version in *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzelde: 597–616. Cambridge: MIT Press, 1997.

Loar, B. (1999). David Chalmers' *The Conscious Mind*. *Philosophy and Phenomenological Research* 59: 464–71.

Loar, B. (2000). Should the Explanatory Gap Perplex Us? *Proceedings of the World Congress of Philosophy*: 99–104. Bowling Green, Ky.: Philosophy Documentation Center.

McGinn, C. (2001). How Not to Solve the Mind-Body Problem. In *Physicalism and Its Discontents*, ed. C. Gillett and B. Loewer: 284–306. Cambridge: Cambridge University Press.

McKinsey, M. (1991). Anti-Individualism and Privileged Access. *Analysis* 51: 9–16.

 [Link](#)

end p.305

Montero, B. (1999). The Body Problem. *Nous* 33: 183–200.

 [Link](#)

Nagel, T. (2002). The Psychophysical Nexus. In his *Concealment and Exposure and Other Essays*: 194–235. Oxford: Oxford University Press. An earlier version appeared in *New Essays on the A Priori*, ed. P. Boghossian and C. Peacocke. Oxford: Clarendon, 2000: 434–72.

Papineau, D. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press.

 [Link](#) 

Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press.

- Perry, J (2004a). *Précis of Knowledge, Possibility and Consciousness*. *Philosophy and Phenomenological Research* 63: 172–82.
- Perry, J. (2004b). Replies. *Philosophy and Phenomenological Research* 63: 207–29.
- Perry, J. (2006) Mary and Max and Jack and Ned. In *Oxford Studies in Metaphysics*, Volume 2, ed. D. Zimmerman, New York: Oxford University Press, 2006: 79–90.
- Putnam, H. (1970). Is Semantics Possible? In *Language, Belief, and Metaphysics*, ed. H. Kiefer and M. Munitz: 50–63. Albany: State University of New York Press. Reprinted in his *Mind, Language, and Reality*: 139–52. Cambridge: Cambridge University Press, 1975.
- Rozemond, M. (1998). *Descartes's Dualism*. Cambridge: Harvard University Press.
- Schiffer, S. (1990). The Mode-of-Presentation Problem. In *Propositional Attitudes*, ed. C. A. Anderson and J. Owens: 249–68. Stanford: CSLI Press.
- Shaffer, J. (1963). Mental Events and the Brain. *Journal of Philosophy* 60: 160–66.
 [Link](#)
- Smart, J. J. C. (1959). Sensations and Brain Processes. *Philosophical Review* 68: 141–56.
 [Link](#)
- Stoljar, D. (2001/2005). Physicalism. In *The Stanford Encyclopedia of Philosophy (Winter 2005 Edition)*, Edward N. Zalta (ed.),
URL = <http://plato.stanford.edu/archives/win2005/entries/physicalism/>
- Sturgeon, S. (1994). The Epistemic View of Subjectivity. *Journal of Philosophy* 91: 221–35.
 [Link](#)
- Van Gulick, R. (1993). Understanding the Phenomenal Mind: Are We All Just Armadillos? In *Consciousness*, ed. M. Davies and G. Humphrey: 137–54. Oxford: Blackwell.
- Van Gulick, R. (2006). Jackson's Change of Mind: Representationalism, *A Priorism* and the Knowledge Argument. In Ian Ravenscroft (ed.) *Minds, Worlds & Conditionals: Essays in Honor of Frank Jackson*.
- White, S. (1986). Curse of the Qualia. *Synthese* 68: 333–68. Reprinted in *The Nature of Consciousness*, ed. N. Block, O. Flanagan, and G. Güzeldere: 695–718. Cambridge: MIT Press, 1997.
 [Link](#)
- White, S. (2006). A Posteriori Identities and the Requirements of Rationality *Oxford Studies in Metaphysics*, Volume 2, ed. D. Zimmerman, New York: Oxford University Press, 2006: 91–102.
- Wiser, M., and Carey, S. (1983). When Heat and Temperature Were One. In *Mental Models*, ed. D. Gentner and A. Stevens: 267–97. Hillsdale, N.J.: Lawrence Erlbaum.
end p.306

thirteen Grasping Phenomenal Properties

Martine Nida-Rümelin

Here I present an argument for property dualism. The argument employs a distinction between *having a concept of a property* and *grasping a property via a concept*. If you grasp a property *P* via a concept *C*, then *C* is a concept of *P*. But the reverse does not hold: you may have a concept of a property without grasping that property via any

concept. If you *grasp* a property, then your cognitive relation to that property is more intimate than if you just have some concept or other of that property. To grasp a property is to understand what having that property essentially consists in.

To have a concept of a property is to have a concept one can use to attribute the property to something. If you have the concept of water, then you can use it to attribute the property of being water to liquids. You then have a concept *of* the property of being water. But you may have the concept of water without knowing that being composed of H_2O is essential for being water—without knowing what having the property of being water consists in. In that case, your concept would not enable you to grasp the property. I will propose an account of grasping properties. It is quite easy to find examples where we do *not* grasp the property at issue. But it might be less obvious that we sometimes do grasp properties via concepts. I think that we sometimes do and that a clear case is provided by our understanding of phenomenal properties via phenomenal concepts.

Grasping Phenomenal Properties

To have a particular phenomenal property is to have an experience with a specific subjective feel. If you have a phenomenal concept of a phenomenal property, then you know what it is to have an experience with that subjective feel. You thereby know what it is to have that property: you grasp the phenomenal property via your phenomenal concept. This idea seems natural. But some work needs to be done to explicate the idea within a theoretical framework that relates concepts and properties.
end p.307

I will call the claim that we can grasp phenomenal properties via phenomenal concepts the thesis of *phenomenal essentialism*. Phenomenal essentialism may seem to trivially imply property dualism. The property dualist denies that to have a phenomenal property consists of having physical features. This denial may appear to follow directly from phenomenal essentialism: if a property has a physical nature, then its nature cannot be grasped via phenomenal concepts. But this reasoning is too quick. We need a number of additional substantial assumptions if we wish to argue for property dualism on the basis of phenomenal essentialism.

The Basic Idea of the Argument

A physicalist who is willing to accept phenomenal essentialism can say (a) that we can grasp phenomenal properties *not only* via phenomenal concepts but also via physical concepts, or (b) that there are physical properties (namely, these phenomenal properties) that cannot be grasped via physical concepts, but which can be grasped via phenomenal concepts.

The following claim excludes (b):

1. Cognitive accessibility of physical properties by physical concepts: every physical property can in principle be grasped via some physical concept.
- To exclude (a), I will defend two further claims, one about grasping properties and another about the cognitive relationship between physical and phenomenal concepts:
2. Cognitive transparency: a person who grasps one and the same property via different concepts can in principle find out without further empirical investigation that the two concepts are necessarily coextensive.
3. Cognitive independence: for every physical concept and every phenomenal concept, it is possible for a rational person with arbitrary physical knowledge to understand both concepts without being in a position to conclude that they are necessarily coextensive.

It follows from 1–3 and phenomenal essentialism that no phenomenal property can be a physical property. Suppose some property *P* is both phenomenal and physical. By phenomenal essentialism and the cognitive accessibility of physical properties by physical concepts, *P* can be grasped by a person via a phenomenal concept *C1* and a physical concept *C2*. By cognitive transparency, that person can see that *C1* and *C2* are necessarily coextensive. But by cognitive independence, this is impossible. This is my main argument, in outline. I will formulate each premise more precisely below. I will also discuss each in detail.

Phenomenal Properties, Phenomenal Concepts, and Phenomenal Beliefs

I will have much to say about what determines the extension of phenomenal and physical concepts. But first I will make some preliminary remarks about phenomenal properties, phenomenal concepts, and phenomenal beliefs.
end p.308

Phenomenal properties are often conceived of as properties of experiences. But here I will assume that phenomenal properties are properties of subjects. This is unusual but, I believe, important for my argument and preferable for independent reasons. This terminological decision seems important to me for the following reason: it is more convincing that a subject can grasp what it is to have a certain property on the basis of having the property than on the basis of having an experience that has the property at issue. In the first case, the subject's access to the property is in a sense more direct than in the second.¹

I will assume that concepts are used in thought, but that different people may have the same concept. I will also assume that concepts are partially individuated by implicit assumptions accepted by those who have the concept. For example, the concept of being watery stuff is different from the concept of being water, because the two concepts are associated with different implicit assumptions about what is essential for having the property at issue.

To understand *phenomenal* concepts, it is helpful to consider beliefs involving phenomenal concepts, or *phenomenal beliefs*. Consider Frank Jackson's famous case of Mary. Mary is raised in a black-and-white room and never has color experiences. One day she leaves the room and looks at the blue sky. When this happens, she learns something new about the color experiences of other people (see Jackson 1982). She acquires the phenomenal belief that those people typically have a certain sort of color experience when they see the blue sky. Thus she takes two steps at once: she acquires the phenomenal concept of having blue experiences; and she forms a correct belief involving this concept.

To see that there are two epistemic steps involved, consider the case of Marianna. Marianna, like Mary, spends her life in a black-and-white environment. Then one day her environment changes radically. The tables, chairs, and so forth are painted in many different colors. However, she sees no bananas, tomatoes, or pictures of landscapes. She sees none of the objects whose colors she knows under previously acquired concepts. While looking at four different slides in sequence (a blue one, a green one, a yellow one, and a red one), she may form the false belief that, with respect to hue, the sky looks to other people the way the red slide looks to her. She has acquired the epistemic capacities to ask new questions and make new mistakes. This is explained by the fact that she has new concepts: phenomenal concepts of phenomenal properties.²

Acquiring phenomenal concepts requires having phenomenal properties oneself. But having or having had a particular phenomenal property is neither necessary nor sufficient for acquiring the phenomenal concept of that particular property. It is not
end p.309

sufficient because a sentient being may experience a particular color without forming the phenomenal concept of the property of having that kind of color experience. A sentient creature has the phenomenal concept of the property of having a particular kind of color experience only if that creature can attribute that property under that concept to another sentient being. It is possible to have a particular color experience without being able to attribute having this kind of experience to another being. Many nonhuman animals may be in this position. The other direction also fails to hold. A person who never had an experience of orange might be able to form the concept of having an experience of orange on the basis of her acquaintance with red and yellow.

The Actual and Counterfactual Extension of Terms and Concepts

To explain what determines a concept's extension, it will help to begin by recounting familiar views on natural kind terms. A lesson of Kripke and Putnam's discussions of natural kind terms may be put as follows: the counterfactual extension of a natural kind term depends on features of the entities that fall under the term in the real world. If what falls under the term "water" in the real world is composed of H_2O , then no liquid in counterfactual circumstances counts as being water unless it has the same chemical composition. The same remark applies to the property term "being water" and to the

concept of water. What falls under the water concept in counterfactual circumstances depends on the chemical structure of the stuff that falls under this concept in the actual world. If the stuff falling under our water concept in the actual world were composed of XYZ, then the counterfactual extension of our water concept would differ: its extension in all counterfactual worlds would be XYZ.

It is quite natural and common to think of the counterfactual extension of a term as being represented by a function that returns for every possible world w those entities that fall under the term in the world w . By “falls under the term in the world w ” I mean this: given the relevant facts about things falling under T in the real world, it is appropriate to apply T to x when talking about the world w . For example, given the chemical facts about the stuff called “water” in the actual world, it is appropriate to apply the term in all counterfactual circumstances to H_2O .

Concepts and linguistic terms may—up to a point³—be treated similarly. I propose the following explication for the counterfactual extension of concepts: x falls under the concept C in counterfactual circumstances w if, and only if, it is appropriate, given the relevant facts about the entities falling under C in the real world, to apply C to x when thinking about the circumstances w . For example,

end p.310

because water is composed of H_2O in the real world, it is inappropriate to apply the water concept to a transparent liquid in the lakes on Earth in an imagined counterfactual world where those lakes instead contain XYZ. Therefore, in that world, XYZ does not fall under the water concept. And it is appropriate to apply the concept to H_2O even in counterfactual circumstances where H_2O lacks the superficial qualities that it actually has—qualities such as how it tastes, for example. Therefore, in that world, H_2O does fall under the water concept.

The scientific nature of the liquids falling under the water concept in the *actual* world determines its *counterfactual* extension. A person who fully masters (understands) the concept knows this, perhaps implicitly. Therefore, if my explication is correct, then full mastery of a concept requires implicit knowledge about how it *should* be used when thinking about counterfactual cases. So full mastery of a concept includes implicit, normative knowledge.

In what sense is this knowledge implicit? Consider someone with no knowledge of chemistry and who has never formed the concept of counterfactual extension. This person cannot conceptualize the content of the item of knowledge at issue. Even so, her modal intuitions might accord with what I have said about the counterfactual extension of the water concept. Suppose she is told about the difference in chemical composition between the liquid in the lakes in the actual world and XYZ. She might then judge that an XYZ world is a world without water.

The item of knowledge is normative because it concerns the way in which we *should* use a given concept in our thought. But *social* norms are irrelevant. It is the content of the concept, not social norms, that determines when it is correct or incorrect to apply it in imagined situations.

Counterfactual Extension, the Nature of Properties, and Grasping Properties

The counterfactual extension of a concept is philosophically interesting for the following reason: there is a conceptual link between the counterfactual extension of a concept and the nature of the property expressed by the concept. To know the nature of a property is to know what things that have the property share *necessarily*, where the modality at issue is so-called metaphysical necessity. For example, knowing what conscious individuals happen to have in common does not suffice for understanding consciousness. Perhaps all and only conscious individuals have eyes. But having eyes is not what being conscious consists in. If we wish to know what being conscious consists in, or the nature of this property, then we need to understand what all and only the conscious individuals in all metaphysically possible worlds have in common. To know the nature of, or grasp, a property is to know what features are necessary and sufficient for having that property across counterfactual circumstances. An individual in counterfactual circumstances has the property *P* expressed by the concept *C* if and only if it falls into the extension of the concept *C*. So, one who grasps a property *P* has a concept *C* of that property and can decide for every counterfactual circumstance whether an arbitrary thing does
end p.311

or does not fall into *C*'s extension, when given all the relevant information about that thing.

Let us assume that water is H_2O and that there is nothing more we can learn about hydrogen, oxygen, or chemical composition such that the extension of “is composed of H_2O ” depends on facts that are still unknown to us. This is to say, in David Chalmers's (2006) terminology, that the concept “is composed of H_2O ” is not Twin-Earthable. It follows that a person who knows that water is composed of H_2O and understands the water concept thereby grasps the property of being water. On the basis of her understanding of the water concept, she knows (perhaps implicitly) that liquids in counterfactual circumstances fall into the extension of this concept just in case their chemical composition falls into the concept's actual extension—that is, just in case these liquids are H_2O .

This example illustrates two points about grasping properties. First, in order to grasp a property you may need empirical knowledge in addition to possessing a concept that expresses the property. Complete understanding of a concept is not in general sufficient for grasping the property expressed by the concept. Second, understanding a concept entails knowing the features that entities falling under the concept in counterfactual circumstances have in common with entities falling under the concept in the actual world. Understanding a concept involves implicit knowledge of what one may call *essentiality conditionals*, such as the following conditional claim: If the water concept actually applies to all and only H_2O , then the water concept applies to all and only H_2O in thoughts about all counterfactual circumstances.

My account of grasping properties has two parts. One concerns the nature of a property: the nature of a property *P* can be understood in terms of the counterfactual extension of

some concept *C* of *P*. The other concerns knowledge of this nature: to know the nature of *P* is to know the counterfactual extension of some concept of *C*.

Many philosophers implicitly accept the first part of the account. To see this, note that discussions of phenomenal properties are typically couched in terms of counterfactual extension. The functionalist, for example, claims that every individual in metaphysically possible circumstances who fulfills a particular causal role associated with having a blue experience *ipso facto* has a blue experience. His claim about what having a blue experience consists in *is* a claim about the counterfactual extension of the concept of having an experience of blue. The functionalist is not just saying that all individuals who have states that play the relevant causal roles have blue experiences. He is not just stating an empirical or nomological regularity. Rather, he is making a claim about all metaphysically possible worlds. More generally, philosophical debates tend to be, in effect, debates about the counterfactual extensions of concepts that express the properties at issue.

Thus, the proposed account of grasping properties merely adds one feature to a widely accepted view. And the second feature is plausible. If the nature of a property can be accounted for in terms of the counterfactual extension of some appropriate concept, then it is natural to suppose that understanding that nature
end p.312

(grasping the property) consists in having the appropriate kind of knowledge of that counterfactual extension.⁴

The Danger of Circularity

But what is it to know the counterfactual extension of a given concept *C*? As a first approximation, we may say the following: knowing *C*'s counterfactual extension consists in having the ability under ideal cognitive conditions to decide correctly for any entity *E* whether it falls under *C*, when given all the relevant information about *E*. But the relevant information about *E* must be given in terms of concepts. It is obvious that we need some restriction with respect to these concepts. Without any such restriction, it is too easy to satisfy the above condition. For example, every person is trivially able to determine the counterfactual extension of the water concept if the information about the relevant entities in counterfactual circumstances is given to her in terms of the water concept itself.

But the triviality problem cannot be solved by simply excluding the concept itself from how the information is described to the epistemic subject. This restriction is at once too strong and too weak. It is too strong because it would exclude the case of grasping phenomenal properties, which is precisely the case that interests us here. If you wish to decide whether a given subject in counterfactual circumstances falls under the phenomenal concept of having a blue experience, then you must be allowed to adopt that person's perspective. This is so because it is the way things appear to a subject that is directly relevant for determining whether it falls into the extension of a given phenomenal concept. Therefore, we cannot acquire the relevant information about the

counterfactual world unless we are allowed to take the perspective of every subject in the counterfactual world under consideration. But the point of taking the perspective of a subject is, of course (in the present context), to be able to think of that subject under the phenomenal concept itself that is at issue. So the simple restriction under consideration is not acceptable; it excludes too many cases.⁵ But the restriction is also too weak. Suppose a person knows for any individual *A* in any counterfactual circumstance whether *A* falls under a given concept *C*. Suppose also that she has this knowledge only when given information about *A* in terms of concepts *C1*, *C2*, ... *Cn* and that these do not allow her to grasp the properties they express. Then we must say that
end p.313

she has not yet grasped the property expressed by *C*. Therefore, the information about *A* must be given in terms of concepts that allow the epistemic subject to grasp the properties they express.

So we have the following situation: to grasp a property via a concept *C* is to be able to decide (under ideal cognitive circumstances) whether an individual *A* in counterfactual circumstances falls under the concept *C* when all the relevant information about *A* is itself given in terms of concepts that allow the person to grasp the property at issue. At this point, a danger of circularity arises. One may object, You defined *grasping a property* in terms of possessing a certain ability, but in explaining that ability, you used the relevant notion of grasping.

The way out of this difficulty is to readily admit that we cannot give a *definition* of grasping a property. My purpose here is *not* to conceptually reduce the notion of grasping properties to anything simpler and better understood than the notion itself. The account explicates the relation between the notion of grasping properties and counterfactual extension of concepts. The account may be illuminating even if the notion of grasping appears on both sides of a biconditional expressing this relation.

The worry about circularity will recur below, when I present a two-dimensional account of grasping properties. According to this account, a concept allows a person to grasp the property it expresses if and only if it is, as I will say, actuality independent. (This means, roughly, that the counterfactual extension does not depend on the world taken as actual.) In the intuitive explanation of this technical account, it will be impossible to avoid using the intuitive notion of grasping properties again. So here we have the following situation: a technical account of the intuitive notion of grasping properties is given within a technical framework, but the technical framework is given its intuitive interpretation by using the intuitive notion of grasping again. So, we explain the intuitive notion of grasping by using the intuitive notion of grasping.

It would be of no help to say that the technical account is doing the whole explanatory work and that the intuitive interpretation of the technical account is only an unsubstantial addition to help one understand the technical account: without its intuitive interpretation, the technical account is nothing but an empty formal apparatus. We have to admit that we cannot give a technical account of grasping properties without presupposing the pretheoretical intuitive notion of grasping properties.

But this is a well-known situation. In logic, we cannot give a technical account of “and” without presupposing the intuitive notion of “and.” Although this is puzzling for

intelligent beginners, there is no reasonable doubt that the technical account of “and” in logic is nonetheless theoretically helpful and illuminating. Therefore, we cannot argue against a given technical account of an intuitive notion *N* merely by pointing out that the intuitive interpretation of the technical account of *N* requires using *N* once again. The analogy between the technical account of “and” in logic and the technical account of grasping properties within the two-dimensional framework is imperfect. The main difference is that the first notion is unproblematic and well understood before we begin to search for a technical account, whereas the latter notion is problematic and cannot be said to be well understood pretheoretically. We need no logics to justify using “and.” But we need a theoretical account of grasping

end p.314

properties to justify its use in philosophy. This difference is relevant for the question of whether the circle at issue is vicious. In the case of “and,” we may say that the circle is not vicious because the technical account is not supposed to clarify the intuitive meaning of “and.” We cannot use this argument in the case of grasping properties. The theoretical and, in particular, the technical account in this case *is* supposed to clarify the content of the pretheoretical notion of grasping. We must therefore say that a technical account can contribute to our intuitive understanding of a given notion even if the account has to be intuitively interpreted using the intuitive notion itself. In my view, however, this possibility obviously exists. A technical account of a notion *N* can clarify conceptual interrelations between *N* and other intuitive notions and can thereby clarify *N* even if the technical account of *N* is circular in the way at issue. This claim, in my view, is well illustrated by the technical account of grasping properties that I will sketch in this chapter.

I can now give a more precise answer to the question asked at the beginning of the present section: an epistemic subject *S* grasps a property via a concept *C* if and only if *S* can in principle (under ideal cognitive conditions) correctly decide whether an individual *A* falls under the concept *C* when *S* is given all the relevant information about *A* in terms of concepts that allow *A* to grasp the properties expressed by these concepts. The case in which the concept *C* is used in the description is not excluded. The account of grasping a property is circular. But since the account is not intended as a definition, the circularity is not vicious.

Essentiality Conditionals

Grasping a property entails having conceptual knowledge and (in general but not always) empirical knowledge. In the general case, a person grasps a property if she has (a) conceptual knowledge that suffices to implicitly know the relevant essentiality conditionals and (b) empirical knowledge that allows her to implicitly conclude which essentiality conditional has a true antecedent. The water example can be used to illustrate this point.⁶ In order to grasp the property of being water *via* her concept of being water

and on the basis of her background knowledge, a person *P* must fulfill the following conditions:

(C1) *P* knows implicitly that the following essentiality conditional is true:

(EC) If those liquids falling under our concept of water in the real world are composed of H_2O , then a liquid in counterfactual circumstances falls under the concept of being water just in case it is also composed of H_2O .

(C2) *P* knows that *this* essentiality conditional (rather than, e.g., the one involving XYZ)

has a true antecedent.

To fulfill (C1) is to have an item of conceptual knowledge, and to fulfill (C2) requires empirical knowledge. As in the present water example, to know which
end p.315

essentiality conditional associated with a concept has a true antecedent normally requires that one have empirical knowledge about the actual extension of the concept at issue. A lesson to be drawn from this simple consideration is that any hypothesis about the nature of a property expressed by a concept *C* relies in general on (1) a conceptual claim about the essentiality conditionals associated with the concept and (2) empirical knowledge about the entities in the actual extension of *C*. To fully grasp a property, it is not enough to fully understand some concept that expresses the property. This is because grasping a property may require empirical knowledge in addition to conceptual knowledge. In general, it is not sufficient for grasping the property that a concept expresses that one fully understand the concept because empirical knowledge is required in addition to conceptual knowledge. However, there are exceptions: some concepts are such that if you understand the concept, you thereby grasp the property it expresses. The present simple model of grasping a property allows us to say how a concept can have this particular status. Having a concept implies grasping the property it expresses if having the concept necessarily involves knowing which essentiality conditional has a true antecedent. I will argue later on that this condition is met by phenomenal concepts.⁷

A Two-Dimensional Account of Grasping Properties

In this section, I will develop a two-dimensional account of grasping properties. Because this section is technical, some readers may wish to skip it.

My account assumes an interpretation of the two-dimensional framework developed by Chalmers and Jackson (see Chalmers and Jackson 2001, and Chalmers 2006). Here is the basic idea of the account. To grasp a property is to have a concept *C* of that property and to know its secondary intension. To know the secondary intension of a concept is to have a concept *C* such that all possible secondary intensions relative to one's background knowledge (those secondary intensions that the concept might have, given what one knows about its actual extension) coincide with *C*'s real secondary intension. Therefore, one grasps a property if one has a concept of that property that is actuality independent in

this sense: a person who has that concept thereby has background knowledge that is compatible with no more than one secondary intension of the concept. I will argue that phenomenal concepts are actuality independent in this sense. To have a phenomenal concept *PC* involves having background knowledge about *PC*'s actual extension that reduces the set of possible secondary intensions to a single secondary intension, the real secondary intension of *PC*. Therefore, to have a phenomenal concept involves grasping the property it expresses.
end p.316

A Two-Dimensional Account of Grasping

The counterfactual extension of a concept may be identified with its secondary intension. To define the secondary intension of a concept, we must introduce a two-dimensional function that characterizes the concept.⁸ The two-dimensional function F_C that characterizes the concept *C* is a two-place function that returns for every pair of possible worlds w_1 and w_2 a set of entities. Intuitively, the function returns the set of entities in the counterfactual world w_2 that would fall into the extension of the concept *C* if w_1 were the actual world. So the claim

$$F_C(w_1, w_2) = E$$

is to be read as follows: the extension of *C* in the world w_2 would be *E* if w_1 were the actual world.⁹

The real counterfactual extension of a given concept (the counterfactual extension of the concept, given the real relevant features of the actual world) is given by the function that returns for every counterfactual world w_2 the extension of the concept *C* if the actual world has the features it really has. So, if we assume that some possible world w_{actual} is the real world, then we may use the two-dimensional function described above to define the real counterfactual extension of a concept: it is the one-place function of possible worlds into extensions that we get if we put w_{actual} into the first slot of the function F_C and keep it fixed. This idea is captured in the following definition of the secondary intension of a concept.

Definition 1: The secondary intension SI_C of a given concept *C* is defined as follows:

For every w , $SI_C(w) = F_C(w_{\text{actual}}, w)$ where w_{actual} is the actual world.

Using this notion, we can reformulate our account of grasping properties as follows: to grasp a property is to know the secondary intension of some concept *C* of that property. In many cases, one's empirical knowledge combined with one's understanding of the concept at issue does not suffice for knowing the secondary intension. In this case, one's available background knowledge is compatible with different secondary intensions of the relevant concept. The background knowledge of a person who does not know the chemical composition of water is compatible with the assumption that the secondary intension of the water concept is the function that returns for every counterfactual world

w the liquids in w that are composed of XYZ. This knowledge is also compatible with the assumption that the secondary
end p.317

intension of this concept is a function that returns for every counterfactual world w the liquids in w that are composed of H_2O .

It is natural to express this last idea by referring to *possible* secondary intensions. Ideally, we may ask for every possible world w what would be the secondary intension of a given concept C if w were the actual world. The function that returns for every counterfactual world w the liquids in w that are composed of XYZ is a possible secondary intension of the concept of water: this function would be the secondary intension of that concept if the actual world were a world with liquids composed of XYZ in the rivers, lakes, and oceans on Earth. To get the possible secondary intension of a concept that would be the real secondary intension if w were the actual world, we have to put w into the first slot of the two-dimensional function F_C and keep it fixed. We thus get what we may call the possible secondary intension of C relative to w , which we may define as follows.

Definition 2: The possible secondary intension of a concept C relative to the possible world w_1 SI_{C,w_1} is defined as follows:

$$w_2: SI_{C,w_1}(w_2) = F_C(w_1, w_2)$$

Earlier I claimed that understanding a concept entails knowing associated essentiality conditionals. On this view, in understanding a concept, you know what features an entity E in counterfactual circumstances must share with the entities falling under the concept in the actual world if E falls into the counterfactual extension of the concept. Understanding the concept of water thus involves knowing that a liquid in counterfactual circumstances falls into the extension of the concept just in case it is composed of H_2O , *if* the real world is a world with liquids composed of H_2O in the rivers, lakes, and oceans on Earth. In other words, if you understand the concept, then you know that the function returning liquids composed of H_2O for every counterfactual world w would be the secondary intension of the concept *if* the real world were one of those possible worlds where there is H_2O in the rivers, lakes, and oceans on Earth. In this case, you know for every possible world w with this particular property what *would* be the secondary intension of the concept of water *if* w were actual. You know the possible secondary intension of the concept of water relative to w . The same is true for every other possible world. If you are given the relevant information about an arbitrarily chosen possible world w (the information about the chemical composition of the liquids in the rivers, lakes, and oceans in w), then you know what would be the secondary intension of the concept of water if w were actual. This intuitive reasoning justifies the following proposal: to know the essentiality conditionals of a given concept C is to know all its possible secondary intensions.

Suppose that your background knowledge, together with your understanding of a concept C , do not put you in a position to know C 's secondary intension. In that case, there are still possible worlds that might for all you know be the actual world and fulfill the following condition: if such a world w were the actual world, then C 's secondary intension would be different from its real secondary intension. Now suppose your

background knowledge and understanding of C does put you in a position to know C 's secondary intension. In that case, any world w that might still
end p.318

be actual according to what you know fulfills this condition: the possible secondary intension relative to w coincides with C 's real secondary intension. We are thus led to the following semiformal definition of grasping properties:

Definition 3: An epistemic subject S grasps the property P iff S has a concept C of P and the set K that represents her background knowledge fulfills the following condition:

$$\forall w (w \models K \rightarrow SI_{Cw} = SI_C)$$

In a case in which a subject fulfills the *definiendum* of definition 3, the counterfactual extension of her concept C does not depend on any features of the world not yet known by someone who has background knowledge K . Given K , the concept is in this sense actuality independent. This motivates the following definition:

Definition 4: A concept is actuality independent relative to background knowledge K iff

$$\forall w (w \models K \rightarrow SI_{Cw} = SI_C)$$

There may be concepts that are actuality independent relative to any background knowledge (I will argue that phenomenal concepts are of this kind). Any person who has such a concept knows its secondary intension. In these cases, whoever has the concept thereby grasps the property it expresses. Concepts of this special kind fulfill the following definition:

Definition 5: A concept is actuality independent iff

$$\forall w SI_{Cw} = SI_C$$

According to this definition, all possible secondary intensions of an actuality-independent concept coincide with its secondary intension. This means that nothing a person who has that concept can learn about entities falling under the concept in the real world will change her judgments about the counterfactual extension of the concept. Also, she need not learn anything about the world (anything in addition to what she knows, given her understanding of the concept) to know its counterfactual extension.

So there are two cases of grasping properties to distinguish. In some cases (if a person has an actuality-independent concept of a property), having a concept suffices for her to grasp the property it expresses. In other cases, additional empirical knowledge is needed. Someone who has an actuality-independent concept of a property thereby trivially has an actuality-independent concept of that property relative to what he or she knows.

Therefore, we can cover both cases of grasping with the following short formulation: to grasp a property is to have an actuality-independent concept of that property relative to one's background knowledge.

There is a close relation between the notion of an actuality-independent concept and the property of having identical primary and secondary intensions: a concept is actuality

independent if and only if its primary intension coincides with any of its possible secondary intensions (for the definition of primary intensions, see below). If we interpret Chalmers's claim that the primary and secondary intensions of phenomenal concepts coincide as the stronger claim that the primary intension of these concepts coincides with any of its *possible* secondary intensions, then the
end p.319

thesis that phenomenal concepts are actuality independent in the sense of definition 5 is equivalent to Chalmers's claim (Chalmers 1996, 2003b).
We still must define the primary intension of a concept.

Definition 6: The primary intension PI_C of a given concept C is defined as follows:

$$\forall w: PI_C(w) = F_C(w, w)$$

So the primary intension of a concept C returns for every possible world w what would be the extension of the concept C in w if w were the real world. Therefore, the primary intension can be interpreted as capturing the way the real extension of a concept is determined. On this interpretation, implicit knowledge of the primary intension is to be understood as knowledge about how the extension of a concept is determined in the actual world.

Knowledge and Sets of Possible Worlds: A Problem for the Proposed Account

The proposed two-dimensional definition of grasping properties presupposes that knowledge can be represented by sets of possible worlds. Within the two-dimensional framework, the relevant set can be determined in two different ways. It can be determined on the basis of primary intensions, thereby representing (according to a well-known but controversial interpretation) the subjective content of the belief or knowledge. The set can also be determined on the basis of secondary intensions. For the purposes of defining possible secondary intensions relative to background knowledge, we cannot represent the relevant knowledge on the basis of secondary intensions. For then we would be unable to distinguish appropriately between a person who knows that water is H_2O (Maria) and a person who has no chemical knowledge at all (Anna). Suppose we chose the set of possible worlds representing their knowledge on the basis of secondary intensions. In that case, the possible secondary intension relative to a given world w of the concept of water would be a function that returns liquids composed of H_2O for every possible world w in the set representing Maria's knowledge as well as for every possible world in the set representing Anna's knowledge. But what Anna knows is compatible with the assumption that the real possible secondary intension of the concept of water returns XYZ for every counterfactual world.

But there are also problems with representing knowledge by sets of possible worlds chosen on the basis of primary intensions. If we accept—as I think we should—that concepts (and thus the subjective content of beliefs involving them) are partially

constituted by the associated essentiality conditionals, then we have to conclude that beliefs may be cognitively different even if they involve concepts with identical primary intensions. For example, beliefs involving the notion of watery stuff are cognitively different from beliefs involving the notion of water, and the relevant cognitive difference is not captured by the observation that the beliefs involve different properties or secondary intensions. Thus, sets of possible worlds chosen on the basis of primary intensions do not capture subjective content.

A further problem concerns the relation between primary intensions and cognitive independence. A close relation between cognitive independence and primary
end p.320

intensions is presupposed in the assumption that we can represent subjective belief content by sets of possible worlds chosen on the basis of primary intensions. The project of representing subjective belief content in this way can succeed only if the following condition is met:

(C) If a rational person who understands the concepts *C1* and *C2* can believe that something falls under *C1* without falling under *C2*, then the primary intensions of *C1* and *C2* differ.

But many physicalists propose accounts of phenomenal concepts and their reference that are incompatible with (C).¹⁰ Let *C2* be a neurophysiological or functional concept of the property thought to be identical with a particular phenomenal property, and let *C1* be the phenomenal concept of that phenomenal property. According to these physicalists, a rational person may have *C1* and *C2* and yet believe that something falls under *C2* without falling under *C1* and vice versa. But they tell a story about how the reference of phenomenal concepts is established that implies that *C1* and *C2* necessarily pick out the same entities in the real world (that is: *C1* and *C2* have the same primary intension). Therefore, to describe belief content in a way that presupposes (C) is to build the denial of some of those physicalist theories that I wish to attack into the conceptual apparatus used in the formulation of the argument. Plainly, this should be avoided.

Conceptualizations: Sketch of a Possible Solution

An alternative is to represent knowledge without using sets of possible worlds. A more neutral way of representing knowledge or belief about the actual world is in terms of what one may call *conceptualizations*. Let us think of descriptions in terms of concepts. Then a conceptualization of the actual world is a description of the actual world that may be incomplete in many respects. To conceptualize the world in a particular way is, roughly, to believe in the truth of a particular description (given in terms of concepts). To conceptualize the actual world in a way given by a certain description *D* is to understand the concepts used in *D* and to believe that *D* is true as a description of the actual world.¹¹

Using the notion of conceptualizations, we may consider an alternative account of grasping properties, given by the following four definitions. We will say that a function is a possible secondary intension of a concept C given a certain conceptualization D (of the actual world) just in case the assumption that it be the

end p.321

real secondary intension is compatible with that conceptualization. We may define this notion as follows.

Definition 7: The function F from possible worlds into extensions is a possible secondary intension of the concept C given the conceptualization D iff a person who understands C and conceptualizes the actual world according to D cannot exclude that F is the secondary intension of C .

If a person can exclude that *all* functions that are not identical with the secondary intension of a given concept C are C 's real secondary intension, then she knows the counterfactual extension of the concept. She then grasps the property expressed by the concept. We get a new definition of actuality independence on the basis of the following claim: the secondary intension does not depend on features of the actual world that are unknown to a subject if what the subject knows allows for just one possible secondary intension. We can call a concept actuality independent *relative to a certain conceptualization* D of the actual world if C 's secondary intension is the only possible secondary intension relative to the conceptualization D . We can call a concept actuality independent *tout court* if the only possible secondary intension of the concept relative to all conceptualizations D is its real secondary intension. These notions of actuality independence accord with our earlier definition of grasping a property: to grasp a property is to have an actuality-independent concept of that property relative to what one knows about the actual world. The only difference is that we individuate knowledge by conceptualizations and not by sets of possible worlds. We thus get the following alternative definitions:

Definition 8: A concept is actuality independent relative to the conceptualization D iff the secondary intension of C is the only possible secondary intension relative to D .

Definition 9: A concept is actuality independent iff it is actuality independent relative to every conceptualization D .

Definition 10: If D is the description that captures what a person P knows about the actual world (P conceptualizes the actual world according to D and this conceptualization constitutes knowledge), then P grasps the property expressed by the concept C iff the concept C is actuality independent relative to D .

This account of grasping avoids the problem of representing knowledge by sets of possible worlds mentioned in the previous subsection. But there is a disadvantage, too: the central notion of actuality independence used in the account is no longer defined

within the two-dimensional framework. Rather, the account employs the epistemic notion of being able to exclude that a certain function is the real secondary intension of a concept, given one's conceptualization of the actual world. Here again, we have to think of the counterfactual worlds as given in terms of concepts that are not actuality-dependent relative to what the epistemic subject knows about the actual world.¹²
end p.322

Sketch of an Argument for Phenomenal Essentialism

Phenomenal concepts classify subjects according to what is subjectively given in experience. The counterfactual extension of a phenomenal concept of a phenomenal property therefore depends on nothing but the subjectively given—the qualitative feel, the phenomenal character shared by those who fall under the concept in the actual world. We have seen that, in general, the counterfactual extension of a concept *C* depends on the features of those entities that fall into *C*'s actual extension. This holds for phenomenal concepts as well. The counterfactual extension of the phenomenal concept of having blue experiences depends on the qualitative character present in the experience of those who fall under the concept in the actual world. We can formulate this idea, in a first approximation, by the following essentiality conditional:

(EC) If *Q* is some kind of hue quality (*Q* could be the hue we call green, yellow, red, or blue), and if *Q* is experienced by those who fall under the concept of having blue experiences in the actual world, then a subject in counterfactual circumstances falls under the extension of that concept just in case that subject has an experience of *Q*.

The essentiality conditionals associated with a concept are characteristic of the concept. It is part of our phenomenal concept of having blue experiences that the counterfactual extension does not depend on anything but the hue experienced by those who fall under the concept in the real world.

This first observation about phenomenal concepts also applies to nonphenomenal concepts of phenomenal properties. While in their black-and-white environments, Mary and Marianna may communicate with others who have seen colors. Mary and Marianna thereby form nonphenomenal concepts of the property of having blue experiences. As in the case of the phenomenal concept of this property, it is part of their nonphenomenal concept of having blue experiences that no subject falls into its counterfactual extension unless that subject experiences the hue experienced by those who fall into the actual extension of the concept. Therefore, although phenomenal concepts are associated with distinctive essentiality conditionals, this particular feature is not yet what makes them phenomenal concepts. They share this feature with some nonphenomenal concepts of phenomenal properties.

A person who has the phenomenal concept of having blue experiences thereby knows—or so we may say in a first approximation—*which* hue is experienced by those who fall into the actual extension of that concept. He or she knows of some particular hue *Q* that *it* is the quality experienced by those who fall under the concept in the actual world. If this

claim is accepted, then we have something like an argument for the thesis of phenomenal essentialism (the claim that having a phenomenal concept implies grasping the property it expresses). You know the counterfactual extension of a concept (and thus grasp the property it expresses) if you know of one of its essentiality conditionals that it has a true antecedent. But if the reasoning just sketched is correct, then having the phenomenal concept of blue experiences implies knowing which essentiality conditional has a true antecedent.

The reasoning just sketched seems to me to explicate an important intuitive reason for accepting phenomenal essentialism. There are two main ideas involved: the first
end p.323

concerns the essentiality conditionals associated with the phenomenal concept at issue, and the second concerns knowledge about the actual extension of the concept. These ideas are sometimes combined into one, as when people say that phenomenal concepts are not natural kind concepts. What they mean is that the counterfactual extension of these concepts does not depend on some hidden scientific nature.

How the Argument Is Question-Begging but Useful

The argument sketched in the preceding section implicitly presupposes its conclusion. It uses the premise that a person who has the phenomenal concept of having blue experiences thereby knows *of a particular hue* that it is the hue experienced by those who fall under the concept in the actual world. This talk of “knowing of a particular hue” must be interpreted as implying a particular conceptualization: the item of knowledge at issue is what I call elsewhere (1996, 1998) phenomenal knowledge. The person must know *phenomenally* that all those subjects falling under the concept of having blue experiences have a blue experience.¹³ But knowing this phenomenally is nothing other than having this item of knowledge *under the phenomenal concept of blue experiences*. The person must know, under her phenomenal concept of having blue experiences, that those who fall into the actual extension of her concept have experiences of this particular kind. But then this item of knowledge can help to determine the secondary intension of the concept (on the basis of the associated essentiality conditional) only if the counterfactual circumstances are again given in terms of the phenomenal concept under consideration. So we get the following intermediate result: a person who has the phenomenal concept of having blue experiences can decide for any individual *A* in counterfactual circumstances whether it falls under the concept of having blue experiences when given the relevant information about *A* in terms of the phenomenal concept of having blue experiences. Does this capacity justify the claim that she knows the secondary intension of the concept and so grasps the property at issue?

As I explained above, the general capacity to decide whether an individual *A* in counterfactual circumstances falls under a concept *C* implies grasping the corresponding property only if the information about *A* is given in terms of concepts that allow the person to grasp the corresponding properties. Using the framework developed above, we

must conclude: the information must be given in terms of concepts that are actuality-dependent relative to the person's background knowledge. So, the capacity at issue implies that the person grasps the property at issue only if the phenomenal concept of having blue experiences is actuality independent relative to her background knowledge. But this is precisely what we wanted to show. So the argument for phenomenal essentialism sketched in the preceding section presupposes phenomenal essentialism.
end p.324

As an argument in the strict sense (in the sense of a tool to convince a potential interlocutor by showing that the desired conclusion follows from assumptions that he or she should be ready to accept), the argument is obviously useless. But it is not useless for other purposes. If phenomenal essentialism is correct, then the “argument” explains within a more general framework of grasping *why* phenomenal essentialism is correct. This explanation is useful in getting a better understanding of the precise content of phenomenal essentialism. It helps us see that, for example, there are two components involved in the claim of phenomenal essentialism: one about knowledge of actual extension, the other about essentiality conditionals. We cannot use the argument to *justify* phenomenal essentialism. But we can use it to explain its truth and explicate its content. In some cases, an interlocutor gets convinced by a claim simply by better understanding its content. Insofar as the argument can convince by explaining the content of the thesis, it can be useful as an argument nonetheless. The argument is still a useful argument even if it is no argument in the ordinary strict sense.

The failure of the above argument may however appear to be more devastating. The above discussion may appear to reveal that the claim of phenomenal essentialism when spelled out appropriately is trivially true and therefore uninteresting. The discussion has shown that phenomenal essentialism amounts to the claim that a person who has a phenomenal concept *C* thereby knows that a being in counterfactual circumstances falls into *C*'s extension when given the information *under the phenomenal concept C* that the being has this kind of experience. But of course we know the counterfactual extension of every concept in this trivial sense.

This argument, however, is mistaken. In the general case, the capacity of knowing the counterfactual extension of a given concept *C* under the concept *C* itself cannot constitute grasping the corresponding property. The situation is different in the case of actuality-independent concepts. In the special case of actuality-independent concepts, knowing the counterfactual extension in terms of the concept itself *is* grasping the property at issue. We cannot *justify* the claim that a person grasps the property corresponding to a concept *C* simply by pointing out that she knows its counterfactual extension under the concept *C* itself. But it is wrong to assume that a person does not grasp the property corresponding to a concept *C* if she knows the counterfactual extension under no concept other than *C* itself.¹⁴

Grasping Physical Properties via Physical Concepts

If phenomenal essentialism is true, then physicalism is false unless it is possible to grasp physical properties via phenomenal concepts. Materialists commonly assume that we have phenomenal concepts of some physical properties. But the claim that we can *grasp* some physical properties via phenomenal concepts is stronger. It implies that we can grasp the *nature* of physical properties by conceptualizing them phenomenally.
end p.325

But the nature of physical properties is physical. So, it should be possible to grasp that nature in physical terminology (in a physical conceptualization). The physical nature of physical properties should in principle be cognitively accessible for an ideal epistemic subject. In other words, it should be possible to grasp every physical property by some physical concept. We might even say that this principle partially explains what we mean by physical properties. Physical properties are such that they can in principle be fully expressed in physical terminology. But for a property to be fully expressed in a terminology is to be captured in a way that allows an epistemic subject who understands the terminology (has the corresponding concepts) to understand what having that property essentially consists in.

There are strong reasons for the claim that I called earlier the cognitive accessibility of physical properties by physical concepts: for every physical property *P*, there is some physical concept *C* such that *P* can be grasped via *C*. Of course, we human beings need not have the concept at this time in history. The claim is that there is some concept appropriately called physical such that if we had that concept, we would be able to grasp the property.

But there are also reasons to deny the claim at issue. Some argue that physical terminology captures only causal relations: that such terminology fails to capture what might be called the intrinsic nature of those entities (particles, forces, fields, etc.) that stand in those relations (see Flanagan 1992, Strawson 1999). It can be argued that, on some low physical level, two different kinds of particles may not be distinguishable by their causal role. We could imagine the situation like this: we refer to two kinds of particles and introduce them by describing a causal relation standing between them. If the causal relation is symmetric, then we might not be able to distinguish between them by reference to their causal role, but still we would have reason to assume that they differ intrinsically. In this situation, any property concept that makes reference to one of these two kinds of entities would express a property that cannot be fully grasped by any physical concept.¹⁵ Even so, I will argue that the physicalist who appeals to a denial of this premise in his defense against the argument is led into a quite problematic kind of physicalism.¹⁶

To grasp a physical property, it is in general necessary to have empirical physical knowledge in addition to understanding a concept that expresses the property. I therefore propose to work with the following formulation of the premise at issue:

Cognitive accessibility of physical properties by physical concepts: For every physical property *P*, there is a physical concept *C* and physical background knowledge *K* such that a person who understands *C* and has *K* grasps *P*.

end p.326

Cognitive Transparency

Given the premises of phenomenal essentialism and the cognitive accessibility of physical properties via physical concepts, the physicalist claim that phenomenal properties are physical properties cannot be true unless it is possible to grasp one and the same property via a phenomenal concept and via some physical concept.

My account allows that one can grasp a property via two different concepts. A person can grasp the property of being water via the concept of water, given her chemical background knowledge about the composition of liquids falling under the concept in the actual world. She can also grasp the property of being water via the concept of being a liquid composed of H_2O . However, she knows that her concepts are necessarily coextensive, without further empirical investigation. The idea of grasping a property implies fully understanding what having the property consists in. Therefore, every aspect of what it is to have the property should be in principle cognitively accessible to the subject. But then grasping the property in two conceptually different ways should necessarily go along with the capacity to realize that one and the same property has been cognitively penetrated. *A fortiori*, it should go along with the capacity to see that the two concepts are necessarily coextensive. Therefore, if someone accepts the notion of grasping a property at all, then he or she should also accept the following principle of cognitive transparency.¹⁷

A principle of cognitive transparency (CT): A person who grasps a property P via two distinct concepts $C1$ and $C2$ is thereby in an epistemic situation in which it is in principle possible for her to rationally judge that $C1$ and $C2$ are necessarily coextensive.

Cognitive Independence

Let us assume that we have a phenomenal concept $C1$ of a given phenomenal property P and that the physicalist claim that P is a physical property is true. Then the first two premises—phenomenal essentialism and the cognitive accessibility of physical properties—imply that there is a physical concept $C2$ such that a person who has the physical concept $C2$ and the phenomenal concept $C1$ and sufficient physical knowledge K grasps the nature of P via $C1$ and via $C2$. If the third premise—cognitive transparency—is true, then we can conclude that there are pairs of a phenomenal concept $C1$ and a physical concept $C2$ such that a person who has both concepts can, with sufficient physical knowledge, rationally judge that $C1$ and $C2$ are necessarily coextensive. The purpose of this section is to introduce a notion of cognitive independence between concepts such that (1) it is plausible to assume that phenomenal concepts and physical concepts are independent in this way, and (2) this assumption excludes the result just obtained. Once this is accomplished, my argument for property dualism is complete. The four premises lead into a
end p.327

contradiction when combined with the physicalist view that phenomenal properties are identical with physical properties.

Many philosophers agree that phenomenal terms and physical terms are cognitively independent. Many claim that this independence explains our puzzlement about consciousness. Some think that this cognitive independence explains why we are tempted to think that phenomenal properties are nonphysical but that the explanation is compatible with physicalism and that recognizing the source of our puzzlement should make the temptation disappear. In this chapter, cognitive independence plays a different role. I wish to show that a weak claim of cognitive independence leads quite naturally to a dualist position when combined with my three other main premises.

The notion of cognitive independence needed for the present purposes concerns modal judgments (judgments about *necessary* coextensionality) and must be relativized to a certain body of knowledge. This is because we wish to exclude that a person can, on the basis of her physical knowledge, rationally judge that a certain phenomenal concept and a certain physical concept are necessarily coextensive. We can define an appropriate notion of cognitive independence as follows.

Definition 11: The concepts *C1* and *C2* are cognitively independent relative to background knowledge *K* iff a rational person who accepts *K* as true in the actual world and understands *C1* and *C2* is not on that basis in an epistemic position to rationally judge that *C1* and *C2* are necessarily coextensive.

Two concepts *C1* and *C2* can of course be cognitively independent relative to *K* even if *K* implies that they are actually coextensive. They are cognitively independent just in case understanding the concepts and having background knowledge *K* does not enable one to rationally exclude that there is a metaphysically possible world where the concepts have different extensions.

The water example provides a case that does *not* fall under the notion of cognitive independence just defined. Let *K* be a body of knowledge containing the information that the two concepts are coextensive in the real world. Someone who has *K* and has the concept of being water and the concept of being H₂O is thereby in a position to rationally conclude that the two concepts are necessarily coextensive.¹⁸ Two
end p.328

concepts may not be cognitively independent given available background knowledge *K* even if they are cognitively independent in another familiar sense: it is still possible for a person to coherently conceptualize how the actual world would have to be for the two concepts to have different extensions in the actual world. Again the water case can illustrate this point. Although the case does not satisfy the *definiens* in definition 11, we can coherently conceptualize circumstances in which they would have different actual extensions.¹⁹

To find a case that satisfies definition 11, we can take the concept of having a pain, the concept of having a brain with its C-fibers firing, and any purely physical knowledge about the functioning of the brain. Many philosophers accept that physical knowledge about the functioning of the brain and our understanding of the two concepts involved does not enable us to rationally conclude that the two concepts involved (the phenomenal and the physiological one) have the same extension in every metaphysically possible

world. At least, those philosophers who accept that there is an explanatory gap and more generally that there is a temptation to become a dualist given the cognitive independence of phenomenal and physical concepts should be ready to admit this claim. If they deny that phenomenal and physical concepts are cognitively independent relative to purely physical knowledge about the brain, then it is difficult to see how they could still explain dualist temptations on the basis of some other claim of cognitive independence. The reason is this: if it were possible to rationally realize that phenomenal concepts are necessarily coextensive with appropriately chosen physical concepts on the basis of physical knowledge alone, then no puzzlement about consciousness should remain once we have that kind of physical knowledge.

The claim at issue about cognitive independence between phenomenal and physical concepts can be formulated as follows.

Cognitive independence (CGI): If $C1$ is a phenomenal concept, and $C2$ is a physical concept, and if K is some arbitrary physical background knowledge, then $C1$ and $C2$ are cognitively independent relative to K (in the sense of definition 11).

Why Cognitive Independence Is a Weak Claim

In the current literature, much discussion focuses on whether a complete physical description, supplemented by indexical information and a “that’s all” clause, a priori implies claims about the phenomenal. The cognitive independence of the phenomenal and the physical is standardly understood in these terms. According to the standard claim of cognitive independence, there is no such a priori implication. This claim differs from CGI. Let CGI* be the claim we get when substituting K with K^* , where K^* is K plus the indexical information and the relevant “that’s all” clause. The logical relation between CGI* and the standard claim at issue is as
end p.329

follows: (a) The standard claim implies CGI*, but (b) CGI* does not imply the standard claim. To see (a) is easy: according to the standard claim, knowing K^* and understanding $C1$ and $C2$ are not sufficient for rationally judging that $C1$ and $C2$ are actually coextensive. It follows that it is also impossible to rationally judge that these concepts are necessarily coextensive. To see that (b) is true, we need only realize that a person may not be able to rationally judge that two concepts are necessarily coextensive on the basis of what she knows about the actual world, even if she is able to rationally judge that they are coextensive in the actual world. So, CGI* can be true in a case in which the standard claim is false. Given this logical relation, the standard claim is at least as problematic as CGI, which is still weaker than CGI*. A philosopher who accepts the standard claim must also accept the claim of cognitive independence here proposed.

Again, CGI is even weaker than CGI*: K does not contain indexical knowledge. Thus CGI is even compatible with the view of those who think that for an appropriately chosen physical concept $C1$ and phenomenal concept $C2$, adding indexical knowledge to the relevant physical knowledge will make it possible to rationally conclude that $C1$ and $C2$

are necessarily coextensive. Also, CGI is compatible with the assumption that a person who knows *K* and has *C1* and *C2* can on that basis rationally conclude that *C1* and *C2* are coextensive in every world in which *K* obtains.²⁰

A number of recent accounts of phenomenal concepts imply that empirical knowledge about the functioning of the brain can give us reason to believe that a particular phenomenal concept picks out a particular physiological type. For example, according to the view proposed by Brian Loar, we will find the physical property that is identical with some given phenomenal property once we have identified the physical property that triggers the phenomenal concept expressing that phenomenal property (1990/97). One might be tempted to think that the proposed principle of cognitive independence is incompatible with any such theory. But the incompatibility is only apparent. We cannot conclude from the observation that the concept *C1* is triggered by instantiations of the property expressed by *C2* alone that *C1* and *C2* express the same property and are therefore necessarily coextensive. Adding that we understand *C1* and *C2* does not suffice either. To rationally get to the conclusion at issue we have to use a *philosophical theory* about how phenomenal concepts refer. For a person who has both concepts, no physical knowledge will suffice to rationally judge that any of these philosophical theories is correct. Philosophical theories do not just follow from empirical theories about the functioning of the brain. Therefore, the proposed claim of cognitive independence does not presuppose or imply the falsity of these accounts of phenomenal concepts.²¹
end p.330

Summary of the Argument

Here again are the four premises of my main argument:

Phenomenal essentialism (PE): If *C* is a phenomenal concept, then a person who has *C* grasps the property expressed by *C* via *C*.

A principle of cognitive transparency (CT): A person who grasps a property *P* via two distinct concepts *C1* and *C2* is in an epistemic situation where she can in principle rationally judge that *C1* and *C2* are necessarily coextensive.

Cognitive independence of physical and phenomenal concepts (CGI): If *C1* is a phenomenal concept and *C2* is a physical concept, then a rational person with arbitrary physical background knowledge who has the concept *C1* and who has the concept *C2* is not in a position where she can in principle rationally judge on that basis alone that *C1* and *C2* are necessarily coextensive.

Premise of cognitive accessibility of physical properties (CA): For every physical property *P* there is a physical concept *C* and physical background knowledge *K* such that a person who understands *C* and has *K* grasps the property *P*.

Let *CI* be some arbitrary phenomenal concept. Then the argument is intended to reject the following assumption:

Assumption (A): The property expressed by the phenomenal concept *CI* is a physical property.

To reject the assumption A is to show that either (a) the property expressed by *CI* is no physical property, or (b) there is no property expressed by the phenomenal concept *CI*. The latter is an eliminativist view about phenomenal properties: it implies that there are no such properties.

The argument goes as follows:

1. The property expressed by *CI* can be grasped via some physical concept *C2* on the basis of some specific physical background knowledge *K*. (from **A** and **CA**)
2. The property expressed by *CI* can be grasped via *CI*. (from **PE**)
3. A person who has *CI* and *C2* and background knowledge *K* is in a position to rationally judge that *CI* and *C2* are necessarily coextensive. (from 1, 2, and **CT**)
4. A person who has *CI* and *C2* and background knowledge *K* is not in a position to rationally judge that *CI* and *C2* are necessarily coextensive. (from 1, 2, and **CGI**)

But 4 contradicts 3.

So, if we accept each of the premises, the materialist assumption A must be rejected. So, to avoid eliminativism we must accept that phenomenal properties are not identical to physical properties.

Phenomenally Revealed Physical Natures

I still owe to the reader an explanation of why denying CA, while accepting the other premises, is not a promising strategy for the physicalist. The main problem is that denying CA does not take the idea of grasping properties seriously enough. A philosopher who defends physicalism by denying CA while retaining the claim of phenomenal essentialism implies that every phenomenal property has a physical end p.331

nature that can be grasped only under a *phenomenal* concept. But to grasp a property is to be cognitively presented with what is essential for having that property. There is no room for any hidden nature behind the property that we are grasping. But there is no physical feature before our mind when we grasp a property via a phenomenal concept. So how could the features we grasp phenomenally be physical nonetheless?

I cannot see how the materialist can answer this question without abandoning the central intuitive point about grasping properties via concepts. If we can grasp physical properties via phenomenal concepts, then the physical differs considerably from how we usually conceived of it. On this view, the physical can reveal its nature by simply being experienced. Some physical properties would then be essentially experiential, and that experiential aspect would be all that it is to have that physical property. That there are such essentially experiential properties is precisely what the property dualist says. The disagreement would then concern only whether such properties can be called physical. The physicalist who chooses this line has to explain what justifies calling them physical despite their special status. In so doing, she would have to show that the difference between her position and property dualism is substantial, rather than merely terminological.

Thus, the physicalist response that combines a denial of CA with an acceptance of phenomenal essentialism seems unstable. A physicalist who denies that phenomenal properties can be grasped via physical concepts should also deny that they can be grasped via phenomenal concepts. Therefore, the reasons for accepting phenomenal essentialism count against the physicalist response at issue.

Intrinsic Natures of Small Physical Entities

Another problem arises when reconsidering what may be the best reason for rejecting CA: the claim that there are microphysical entities with cognitively inaccessible natures. If this is why we reject CA, then there is an explanation of its falsity: in conceptualizing physical properties, we make reference to microphysical entities that have a physical nature that we are unable to grasp. But if the nature of microphysical entities is the only obstacle to grasping physical properties, then all physical properties that cannot be grasped via physical concepts must be cognitively inaccessible precisely for this and no other reason. In particular, we have to conclude that phenomenal properties cannot be grasped physically because their nature is partially constituted by microphysical entities with hidden intrinsic natures. In other words, the hidden intrinsic nature of microphysical entities is responsible for the fact that phenomenal properties cannot be grasped via physical concepts.

But this result again is quite amazing if we accept phenomenal essentialism. We arrive at the conclusion that (a) we can grasp these physical properties phenomenally, and (b) we cannot grasp them via physical concepts *because* these properties involve microphysical entities with some hidden intrinsic nature. But this implies that we somehow circumvent the difficulty of accessing the nature of microphysical entities by phenomenally conceptualizing the relevant properties. But how could

end p.332

that be? We are not aware of any features attributed to microphysical entities when conceptualizing phenomenal properties via phenomenal concepts. The conjunction of (a) and (b) seems to imply that we grasp the intrinsic nature of microphysical entities when grasping phenomenal properties via phenomenal concepts. But there is nothing of all this

cognitively present in our understanding of what phenomenal properties consist in when we think of them in terms of phenomenal concepts. If a property is essentially constituted by some role of microphysical entities with intrinsic natures hidden from physical conceptualizations, then in grasping it phenomenally we should be aware of the role these microphysical entities play in the constitution of the property and of their intrinsic features. Nothing of all this is going on in the simple case where we understand what having a blue experience essentially consists in on the basis of having a blue experience. Therefore, a physicalist who denies CA and who believes that this denial is justified by the idea of hidden intrinsic natures of microphysical properties must also deny phenomenal essentialism.

Comparison to Other Arguments and Concluding Remarks

My argument for property dualism is similar to well-known arguments by Chalmers and Kripke. In this section, I will compare and contrast my argument with their arguments (with which I will assume familiarity).

Chalmers's argument makes substantial use of primary intensions and of the idea that primary intensions represent what is a priori known by a person who understands the expressions at issue. By contrast, my argument does not make any substantial use of primary intensions, and I do not assume that primary intensions represent a priori knowledge. Unlike Chalmers, I do not assume that cognitively independent concepts (in the sense I have explained) have different primary intensions. Therefore, my argument is not undermined by a number of arguments concerning what a competent speaker knows a priori. In particular, it is not undermined by arguments developed in Block and Stalnaker (1999) against the views of Chalmers and Jackson (2001). The important item of a priori knowledge that plays a substantial role in the present argument is implicit knowledge of essentiality conditionals (or implicit knowledge of possible secondary intensions).

A related difference concerns the resulting diagnosis of the lack of analogy between the water/H₂O-case and the case of consciousness. According to Chalmers and Jackson, facts about water can be deduced from the microphysical facts, whereas facts about consciousness cannot. Some authors doubt the first part of the latter claim, but my argument does not rely on it. My argument is compatible with the view that the term "water" is not a priori related to the microphysical language in the way necessary for an a priori deduction of facts about water from the microphysical facts. My argument locates the relevant difference in another place. The difference concerns modal knowledge (knowledge about what is possible) given a certain amount of background knowledge. To know that water is composed of H₂O suffices to exclude that there is a possible world where water is XYZ (whereas to know that *B* is the physiological-functional basis of blue experiences

end p.333

in the actual world does not suffice, even when supplemented with arbitrary physical knowledge, to exclude that there is a counterfactual possibility where *B* occurs without

blue experiences). This claim can be defended without assuming that we can redefine “water” such that “water is composed of H_2O ” is deducible from the microphysical facts, and it can be defended without assuming the weaker claim that the microphysical facts imply a priori that water is H_2O .²²

Phenomenal essentialism was inspired by Chalmers's (2003a) claim about the identity of primary and secondary intensions in the case of phenomenal terms. When formulated within the two-dimensional framework, the two claims are equivalent. Chalmers's claim, however, does not seem to be intended as a claim about the cognitive capacity of grasping properties and is not imbedded in any theoretical account of grasping properties. By contrast, this interpretation is central to my argument.

Phenomenal essentialism is also related to Kripke's claim that “pain” and other phenomenal terms pick out their referents by noncontingent reference fixers. On his view, pain is picked out by the way it feels, which is essential to pain: a state that feels like pain is necessarily pain, and a pain necessarily feels that particular way. I have employed this intuitively appealing idea within a different theoretical framework. Kripke bases his account on the introduction of an additional property of the state (or of the property) of being in pain, namely, the property of being painful. By contrast, the present account need not assume these additional qualitative properties. The present proposal does not need talk of the painfulness of pain in addition to the property of being in pain.

Furthermore, in contrast to the Kripkean formulation (and to the corresponding claim in the work of Chalmers), the present account of phenomenal essentialism places the intuition in question in the context of a general account of grasping properties and understanding concepts. This makes it possible to distinguish two components of the intuitive idea: (a) phenomenal concepts are associated with special kinds of essentiality conditionals, and (b) having a phenomenal concept involves knowing the truth of the antecedent of one of these essentiality conditionals. These two components cannot be distinguished in common formulations of the intuitive idea under consideration.

Kripke does not explicitly formulate anything like what I call the premise of cognitive accessibility of physical properties. There is, however, a related thesis present in Chalmers's discussion of what he calls type F-monism. In his discussion of whether there is a world that satisfies “ P and not- Q ” (where P stands for the conjunction of all physical truths about our world, and “not- Q ” for “there is no consciousness”), he considers the possibility that a world w verifies but does not satisfy P . He writes in this context:

If a world satisfies P , it must have at least the *structure* of the actual physical world. The only reason why W might not satisfy P is that it lacks the intrinsic properties underlying this structure in the actual world. (On this view, the primary intension of a physical concept picks out whatever property plays a certain role in a given world, and the

end p.334

secondary intension picks out the actual intrinsic property across all worlds.) (2003a: 256–57).

The difference between primary and secondary intensions for physical concepts considered in this passage may be read as the claim that we are in principle unable to form physical concepts of these intrinsic properties with identical primary and secondary intensions; and this may be interpreted as implying that we are in principle unable to

grasp the intrinsic nature of these properties. Under this interpretation, the view considered by Chalmers in this passage is the negation of the premise of cognitive accessibility of physical properties.

Some premise of cognitive independence between physical and phenomenal concepts is used in every contemporary antimaterialist argument. In Kripke (1972), the premise is implied by his claim about the conceivability of a case where there is C-fiber firing going on in someone's brain without the person being in pain (and vice versa). He does not formulate his claim in terms of cognitive independence, and he does not offer a general account of the kind of cognitive independence at issue. Plausibly, his claim could be reformulated in terms of cognitive independence relative to arbitrary physical knowledge in the sense of definition 11, since (a) it is the conceivability of a counterfactual case that is at issue, and (b) it is implicit in the discussion that the case remains conceivable even if arbitrary physical knowledge about the actual world is added. Chalmers and Jackson think of cognitive independence in terms of a priori entailment from the physical facts (together with a "that's all" clause and indexical knowledge). As I have explained, the claim of cognitive independence used in the present argument is weaker than the claim discussed by Chalmers and Jackson (2001).

Chalmers and Jackson agree that there are in general two properties related to a property concept, the one corresponding to the primary and the other corresponding to the secondary intension. Anyone who accepts this view will say that there are two equally acceptable answers to the question of which property is expressed by a given property concept. By contrast, I propose to identify the property expressed by a property concept with the property that corresponds to the secondary intension. This difference relates to an issue about what we conceive of when we conceptualize a particular case. Jackson and Chalmers argue that there are potentially two sets of possible worlds corresponding to a given conceptualization: the set of possible worlds where the thought at issue is true according to primary intensions, and the set of possible worlds where the thought is true according to secondary intensions. They concede that even in a case of a coherent conceptualization, the second set might be empty, but they insist that the first will not be empty. Primary intensions thus guarantee that there is a real metaphysical possibility (the possibility picked out by the primary intension of the relevant sentence) corresponding to every coherent conceptualization. In this conceptual framework, the bridge between conceivability and metaphysical possibility is built in terms of primary intensions. My proposal, however, does not rely on these claims about primary intensions. I build the bridge between conceivability and metaphysical possibility in a different way.

end p.335

Perhaps the most significant distinctive feature of my argument is the explicit formulation and use of the principle of cognitive transparency.²³ This principle is the bridge between understanding concepts and grasping properties and between mere conceivability and real possibility. The principle states that we can see (intuit) necessary connections between properties in case we have concepts that allow us to grasp what is essential for or constitutive of these properties. The principle reformulates a traditional idea that has been quite unpopular among many philosophers for many years, since it is contrary to what has been called the linguistic turn. On my principle, a priori knowledge is not a matter of

linguistic knowledge but rather a matter of grasping relations of necessity that hold between properties independently of our linguistic conventions and of our conceptual capacities.²⁴ The principle is not explicitly mentioned in the arguments by Kripke and Chalmers, nor is it mentioned in the Knowledge Argument. I suspect that the principle is nonetheless implicit in each of these arguments. To show this, however, it would be necessary to argue that something like my cognitive transparency principle must be used in a rigorous reconstruction of these arguments. I leave this task for a different occasion.

Acknowledgments

I had several opportunities to present earlier versions of this argument. Each time I learned a lot. I am grateful to the participants in the discussion in Fribourg, November 2001, in Saarbrücken in June 2002, in Konstanz in July 2002, in Santa Cruz in August 2002, in Lausanne in January 2003, in Gainesville in October 2003, and at Rutgers in October 2003. Several people have given me helpful, detailed comments, both orally and in writing. Substantial changes were motivated by comments by Ana Maria Andrei, David Chalmers, Manfred Kupfer, Joe Levine, and Kirk Ludwig. I am grateful to Max Drömmmer for numerous helpful discussions.

References

- Bealer, G. (1998). Intuition and the Autonomy of Philosophy. In *Rethinking Intuition: The Psychology of Intuition in Philosophical Inquiry*, ed. M. DePaul and W. Ramsey: 201–39. Lanham, Md.: Rowman and Littlefield.
- Block, N., and Stalnaker, B. (1999). Dualism, Conceptual Analysis and the Explanatory Gap. *Philosophical Review* 108: 1–46.  [Link](#)
- Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York: Oxford University Press.
- Chalmers, D. J. (2003a). Consciousness and Its Place in Nature. In *The Blackwell Guide to the Philosophy of Mind*, ed. P. Stich and T. Warfield. Oxford: Blackwell. Reprinted in end p.336
- The Philosophy of Mind: Classical and Contemporary Readings*, ed. D. Chalmers: 247–72. New York: Oxford University Press, 2002.
- Chalmers, D. J. (2003b). The Content and Epistemology of Phenomenal Belief. In *Consciousness: New Philosophical Perspectives*, ed. Q. Smith and A. Jokic: 220–72. Oxford: Oxford University Press.
- Chalmers, D. J. (2006). The Foundations of 2D Semantics. In *Two-Dimensional Semantics: Foundations and Applications*, ed. M. Garcia-Carpintero and J. Macia. New York: Oxford University Press. An abridged version of this chapter, Epistemic Two-Dimensional Semantics, is in *Philosophical Studies* 118, 1–2 (2004): 153–226.
- Chalmers, D. J., and Jackson, F. (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110: 315–61.

Chisholm, R. (1966). *Theory of Knowledge*. Englewood Cliffs, N.J.: Prentice-Hall.

Flanagan, O. (1992). *Consciousness Reconsidered*. Cambridge: MIT Press.

Haas-Spohn, U. (1995). *Versteckte Indexikalität und subjektive Bedeutung*. Berlin: Akademie-Verlag. English translation available at: <http://vivaldi.sfs.nphil.uni-tuebingen.de/Alumni/Dissertationen/ullidiss/index.html>)

Haas-Spohn, U., and Spohn, W. (2001). Concepts Are Beliefs about Essences. In *Gottlob Frege: Philosophy of Logic, Language and Knowledge*, ed. R. Stuhlmann-Laeisz, A. Newen, and U. Nortmann: 287–316. Stanford, Calif.: CSLI Publications.

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly* 32: 127–36.

 [Link](#)

Jackson, F. (2005). The Case for A Priori Physicalism. In *Philosophy –Science—Scientific Philosophy. Main Lectures and Colloquia of GAP.5, Fifth International Congress of the Society for Analytical Philosophy*, Bielefeld, 22–26 September 2003, ed. C. Nimtz and A. Beckermann: 251–65. Paderborn, Germany: Mentis. Available at <http://consciousness.anu.edu.au/jackson/aprioriphysicalism.pdf>

Kripke, S. (1972). Naming and Necessity. In *The Semantics of Natural Language*, ed. G. Harman and D. Davidson. Dordrecht: Reidel. Reprinted as *Naming and Necessity*. Cambridge: Harvard University Press, 1980.

Levine, J. (2001). *Purple Haze: The Puzzle of Consciousness*. Oxford: Oxford University Press.

 [Link](#) 

Loar, B. (1990/97). Phenomenal States. *Philosophical Perspectives 4: Action Theory and Philosophy of Mind*, ed. J. Tomberlin: 81–108. Atascadero, Calif.: Ridgeview, 1990. Revised version in *The Nature of Consciousness*, ed. by N. Block, O. Flanagan, and G. Güzeldere: 597–616. Cambridge: MIT Press, 1997.

Nida-Rümelin, M. (1996). What Mary Couldn't Know: Belief about Phenomenal States. In *Conscious Experience*, ed. T. Metzinger: 219–42. Paderborn, Germany: Schöningh/Imprint Academic.

Nida-Rümelin, M. (1997). The Character of Color Terms: A Materialist View. In *Direct Reference, Indexicality, and Propositional Attitudes*, ed. W. Künne, A. Newen, and M. Anduschus. Stanford, Calif.: CSLI Publications: 381–402.

Nida-Rümelin, M. (1998). On Belief about Experiences: An Epistemological Distinction Applied to the Knowledge Argument. *Philosophy and Phenomenological Research* 58: 51–73.

Nida-Rümelin, M. (2004). Phenomenal Essentialism. In *Perception and Reality: From Descartes to the Present*, ed. R. Schumacher. Paderborn, Germany: Mentis: 332–44.

Nida-Rümelin, M. (forthcoming). *Thoughts about Experiences*. Oxford: Oxford University Press.

Papineau, Davi. (2002). *Thinking about Consciousness*. Oxford: Oxford University Press.

 [Link](#) 

end p.337

Perry, J. (2001). *Knowledge, Possibility, and Consciousness*. Cambridge: MIT Press.

Spohn, W. (1997). The Character of Color Terms: A Materialist View. In *Direct Reference, Indexicality, and Propositional Attitudes*, ed. W. Künne, A. Newen, and M. Anduschus. Stanford, Calif.: CSLI Publications: 351–79.

Stalnaker, R. (2001). On Considering a Possible World as Actual. *Proceedings of the Aristotelian Society*, suppl. volume, 65: 141–56.

Strawson, G. (1999). Realist Materialist Monism. In *Towards a Theory of Consciousness III*, ed. S. Hameroff, A. Kaszniak, and D. Chalmers. Cambridge: MIT: 23–32.

White, S. (1999). Why the Property Dualism Argument Will Not Go Away. Unpublished. Available at:
<http://www.nyu.edu/gsas/dept/philo/courses/consciousness/papers/WHYPDAW.html>